NY Property Fraud Analysis

Prepared by:

Team 4

Wenzhen Zhao, Ying Liu, Ting Gu, Xuan Zhang, Po-Nien Chiang, Yue Shi, Xingjian Zheng

Instructor: Professor Coggeshall

Fraud Analytics (DSO 562)

Table of Contents

❖ Executive Summary

The report provides fraud analysis of New York City Properties using unsupervised algorithms. Main data processing tools are Python, R and Microsoft Power BI.

The original dataset contains 1,048,575 records. The pipeline of our works includes data cleaning, variable construction, zero-mean normalization and dimensionality reduction, fraud algorithm implementation, score calculation and potential fraudulent records identification.

Detailed examination on the high scored properties provides a lot of thought-provoking information, including descriptive analysis and some more detailed study about potential exempt fraud records. Further examination of top 10 highest scored records is delivered one by one according to their addresses, owners and some other special characteristics of their own.

❖ Part I: Data Overview

1.1 Data summary

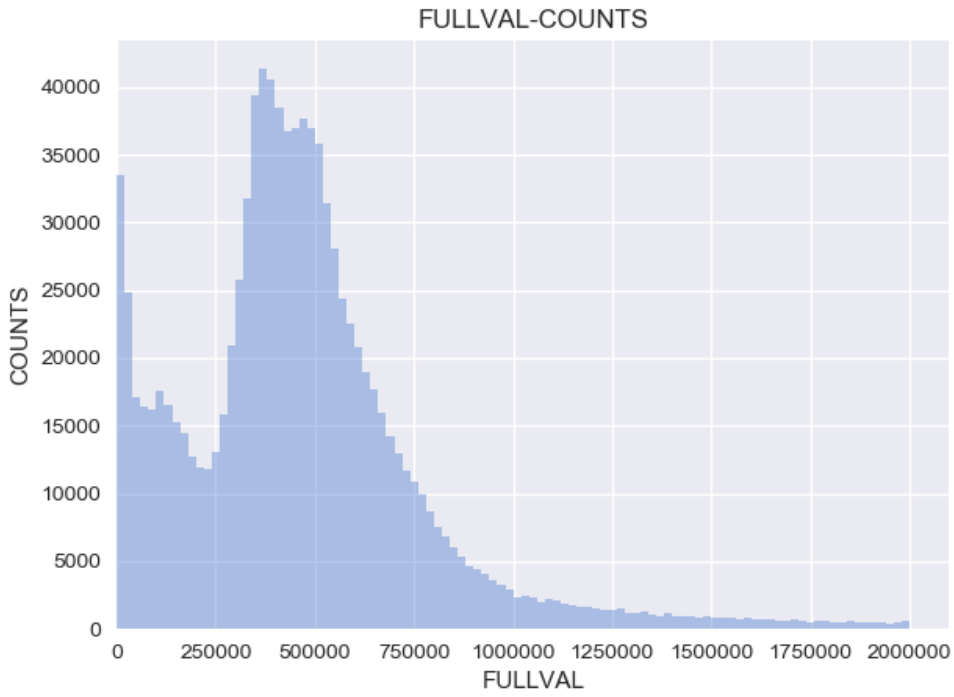
The City of New York Property Data is published on NYC Open Data website (https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8) by the New York City Department of Finance. The dataset contains 1,048,575 records and 30 variables which related to properties' address, value, owner, size, easement, building classes, tax classes and so on. Among these 30 variables, 16 are numeric, 13 are categorical and a date variable.

Below is an excerpt of the variables. The complete Data Quality Report is attached in appendix.

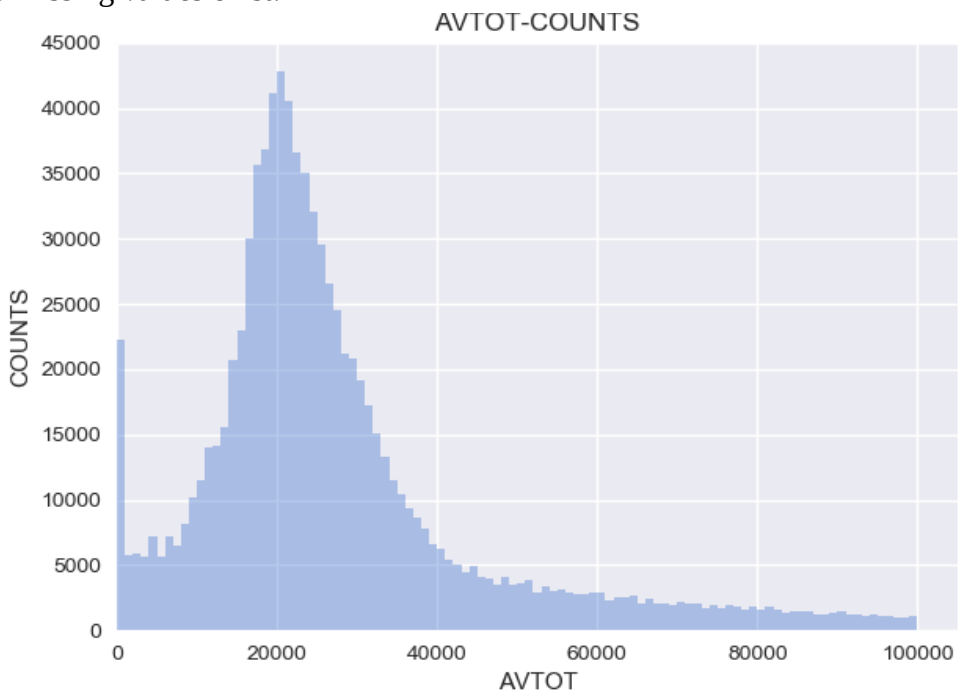
1.2 Important variables (Key variables description)

1.2.1 Base variables

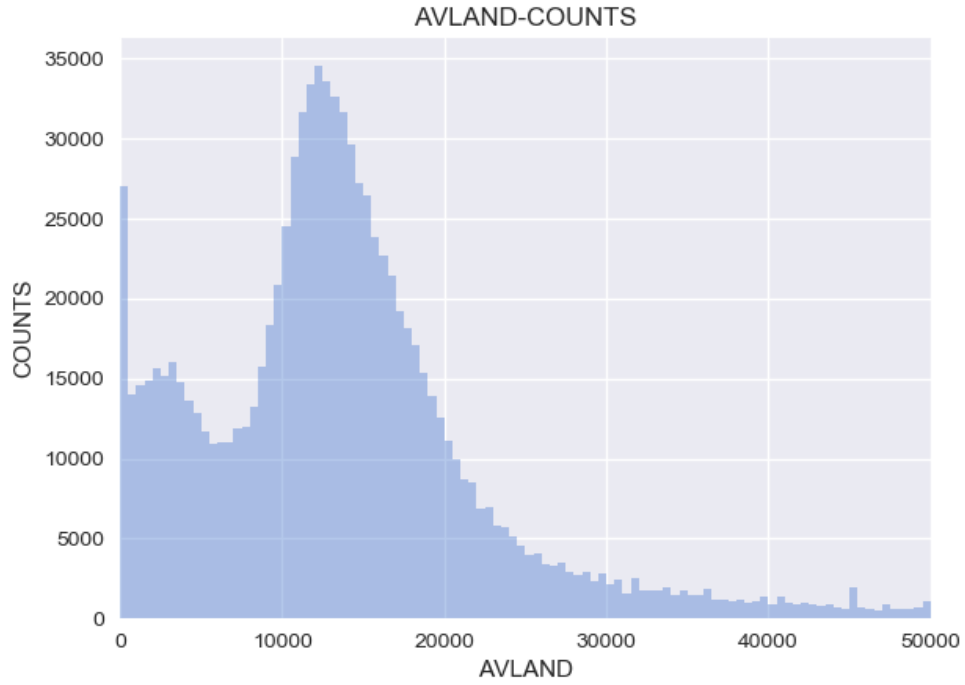
FULLVAL (numeric variable): The full value of the property. There are 108277 unique values ranging from 0 to about 6,000,000,000. There are 12,762 properties with FULLVAL 0 in the dataset. No missing values exist.



AVTOT (numeric variable): The assessed total value of the property. There are 112294 unique values ranging from 0 to about 4,700,000,000. There are 12,762 properties with AVTOT 0 in the dataset. No missing values exist.



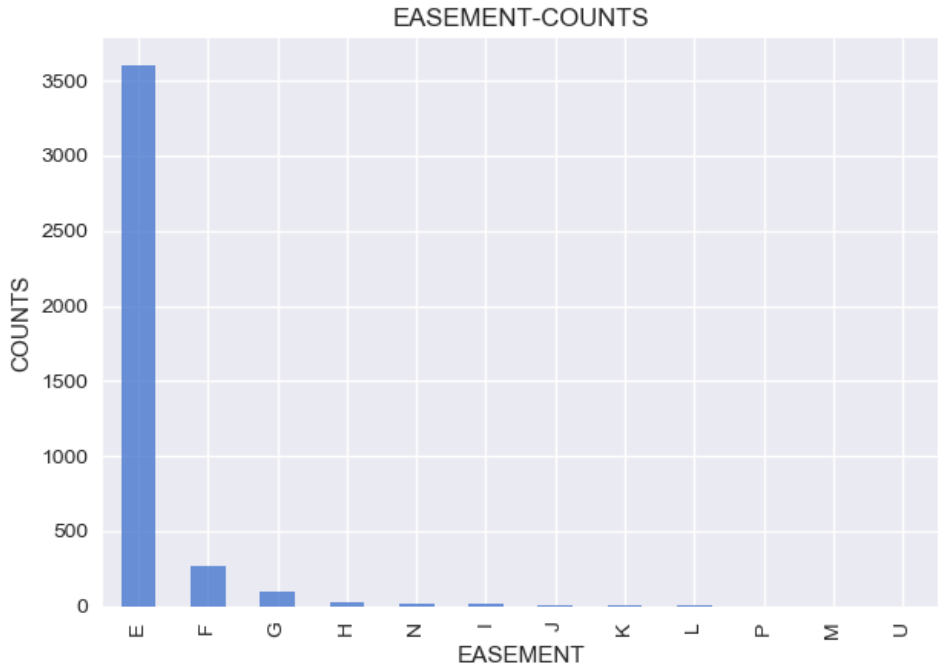
AVLAND(numeric variable): the assessed value of the land. There are 70,529 unique values ranging from 0 to about 2,700,000,000. There are 12,764 properties with AVLAND 0 in the dataset. No missing values exist.



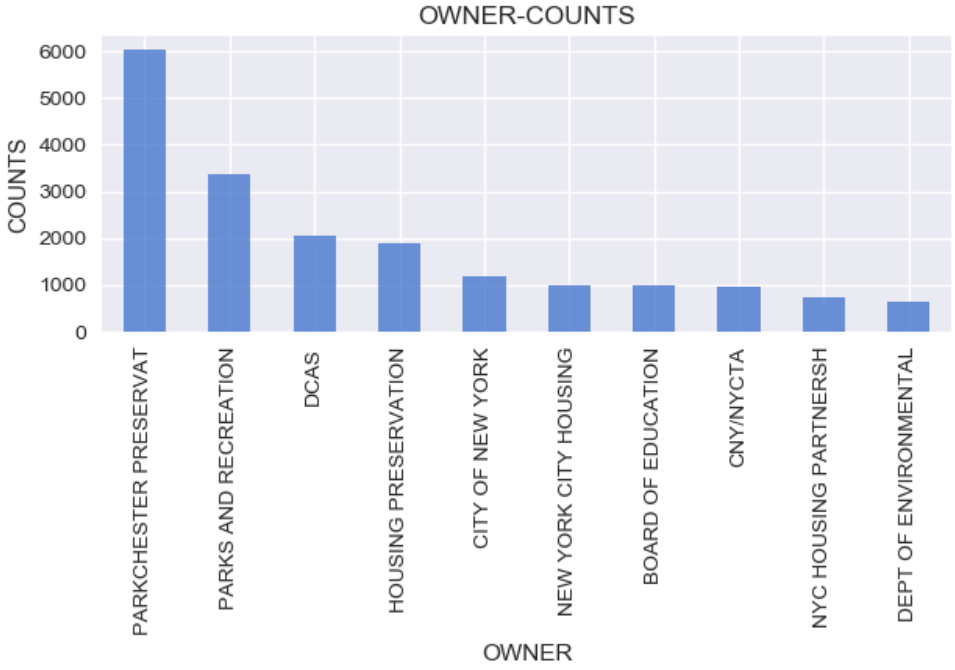
1.2.2 Other key variables

BBLE (categorical variable) : The concatenation of BORO code, BLOCK code (5 digit), LOT code (4 digit) and EASEMENT code. There are 1,048,575 unique values and no repeated values or missing values.

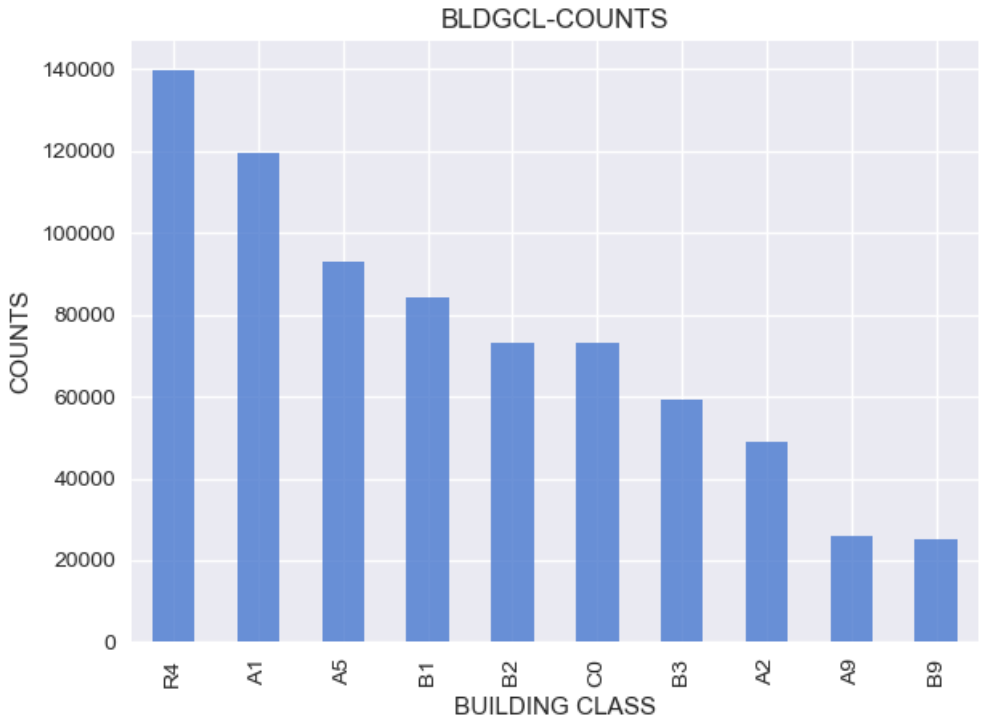
EASEMENT (categorical variable): The property's easement type. There are 13 levels – "", "E", "F", "G", "H", "I", "J", "K", "L", "M", "N", "P", "U". Null value indicates the property does not have any special easement. No missing values exist.



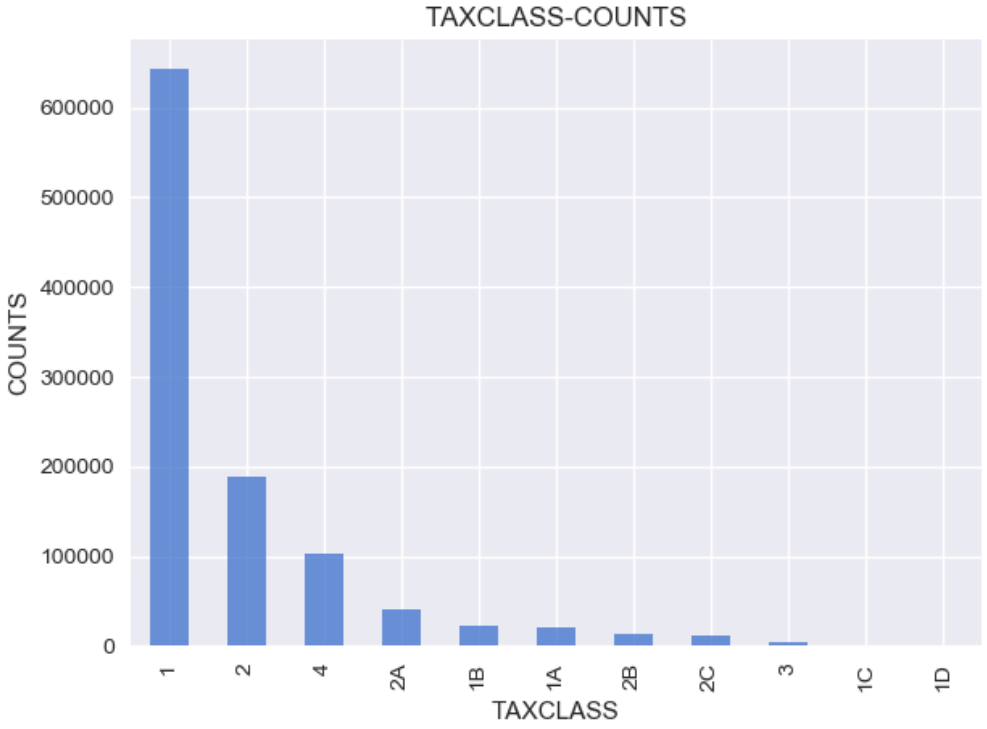
OWNER (categorical variable): The owner of the property. There are 847055 unique values and 31081 missing values.



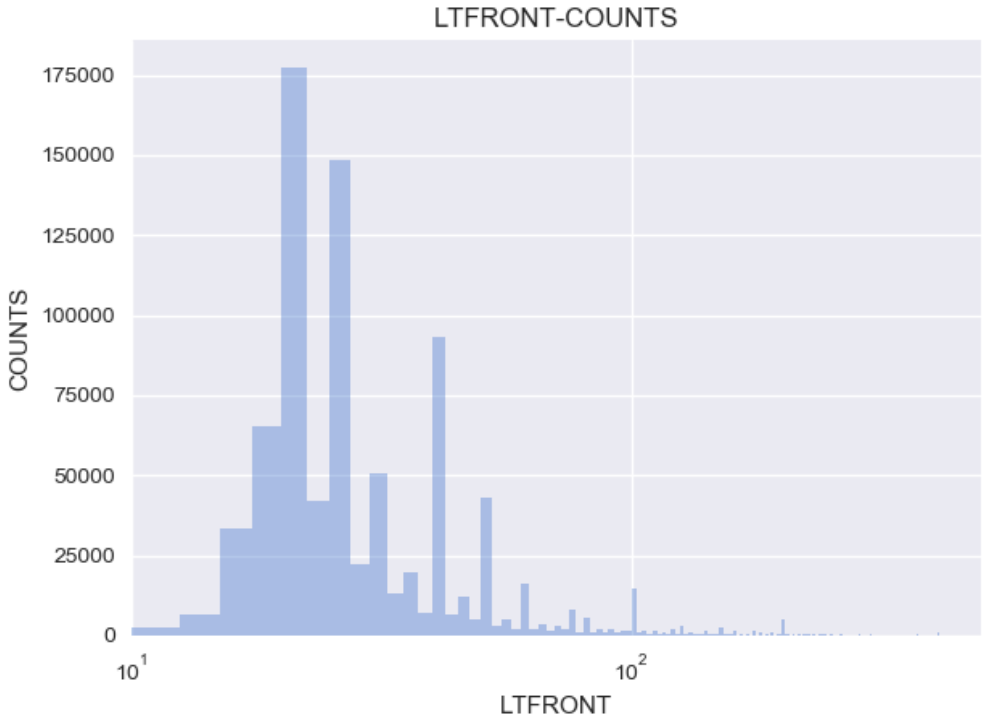
BLDGCL (categorical variable) : Building class. There are 200 unique values and no missing values.



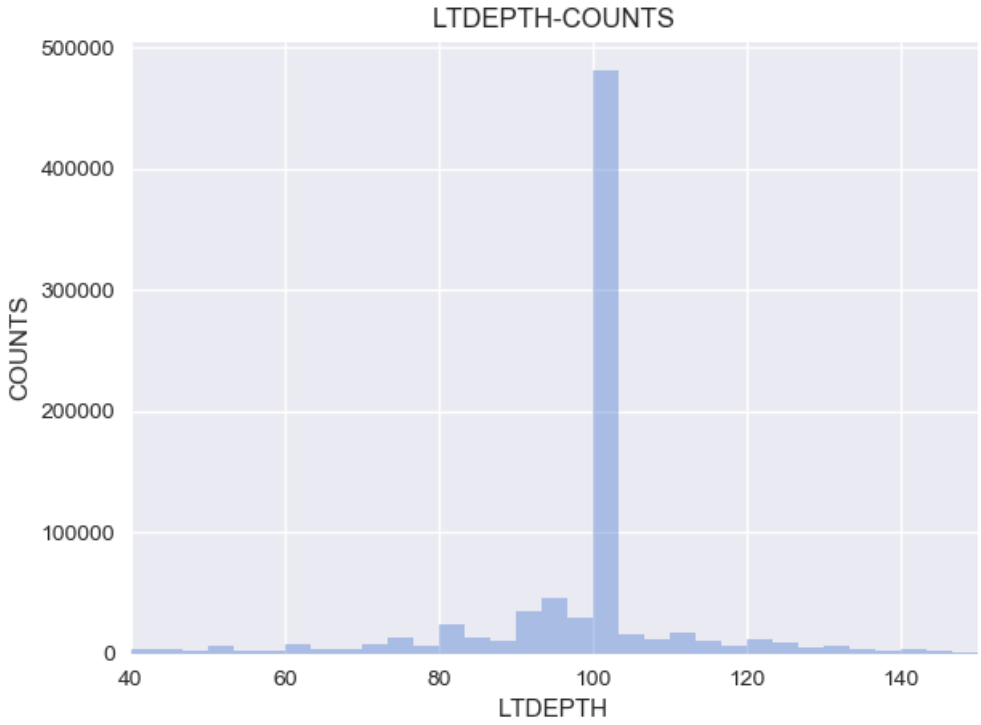
TAXCLASS (categorical variable): The tax class of property. There are 11 unique values and no missing values.



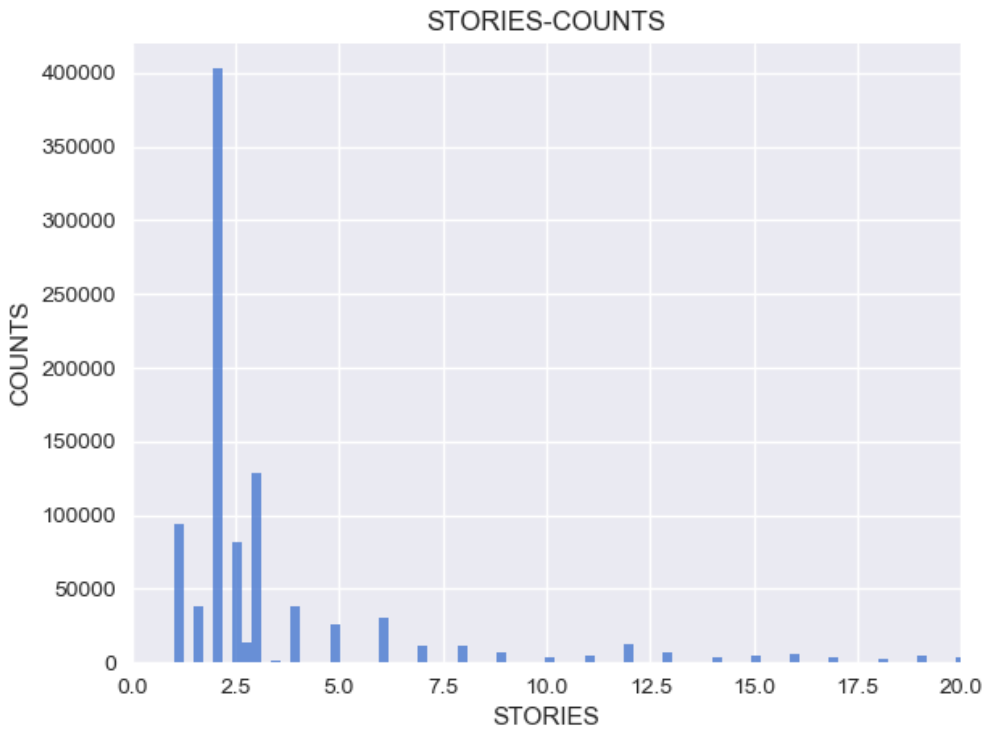
LTFRONT (numerical variable): The length of lot frontage in feet. There are 1277 unique values and no missing values.



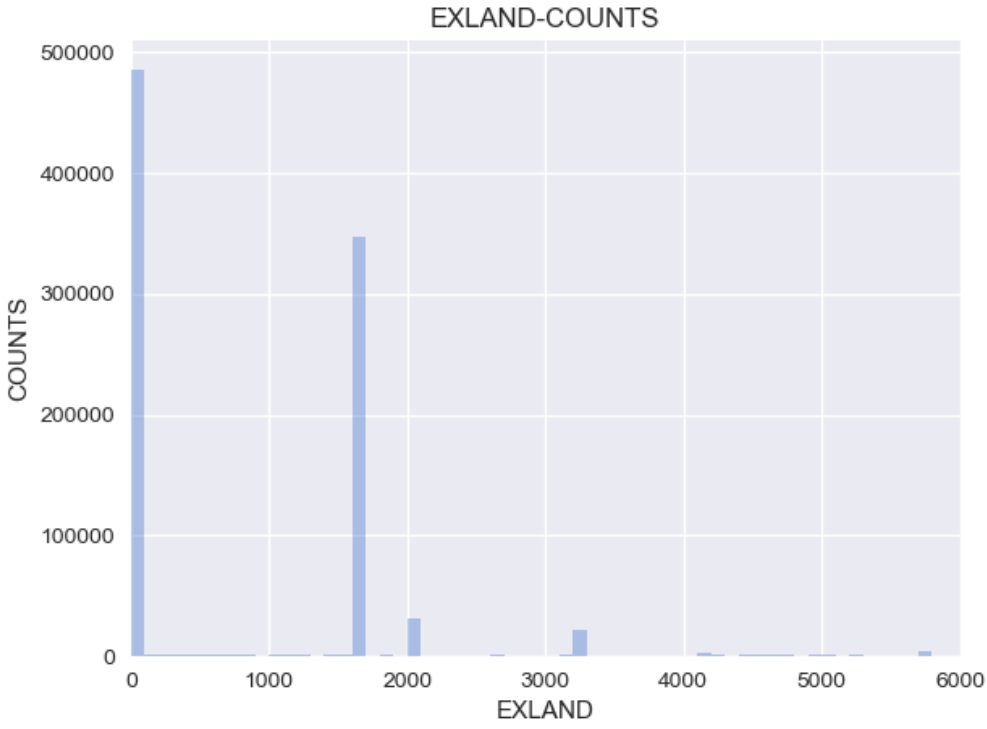
LTDEPTH (numerical variable): The length of lot depth in feet. There are 1336 unique values and no missing values.



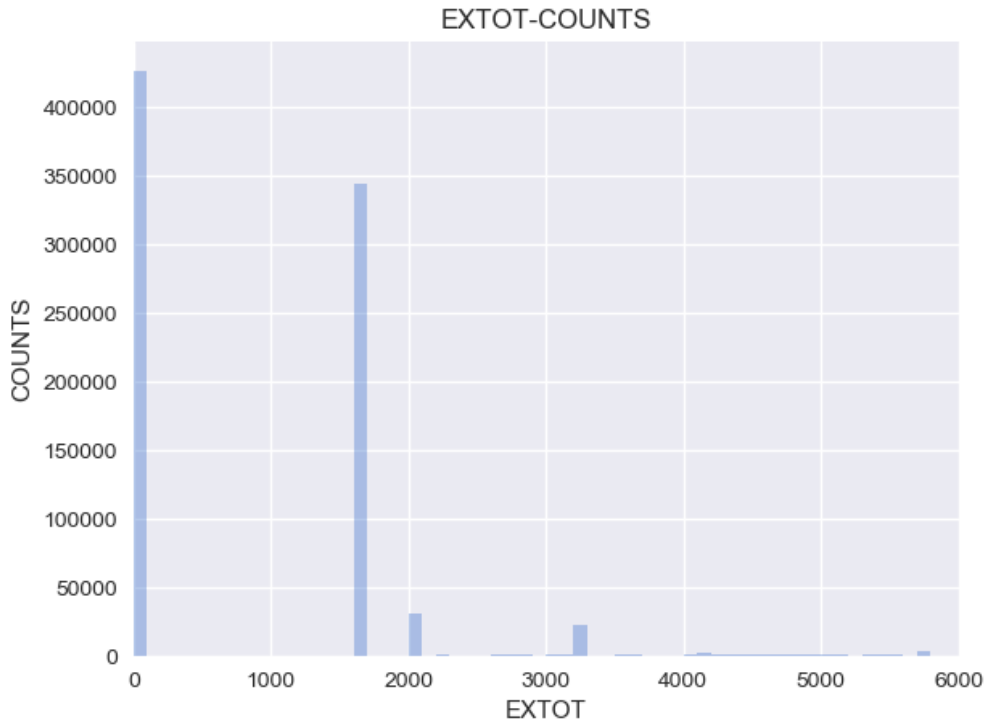
STORIES (numerical variable): The number of stories for the building. There are 112 unique values and 52142 missing values.



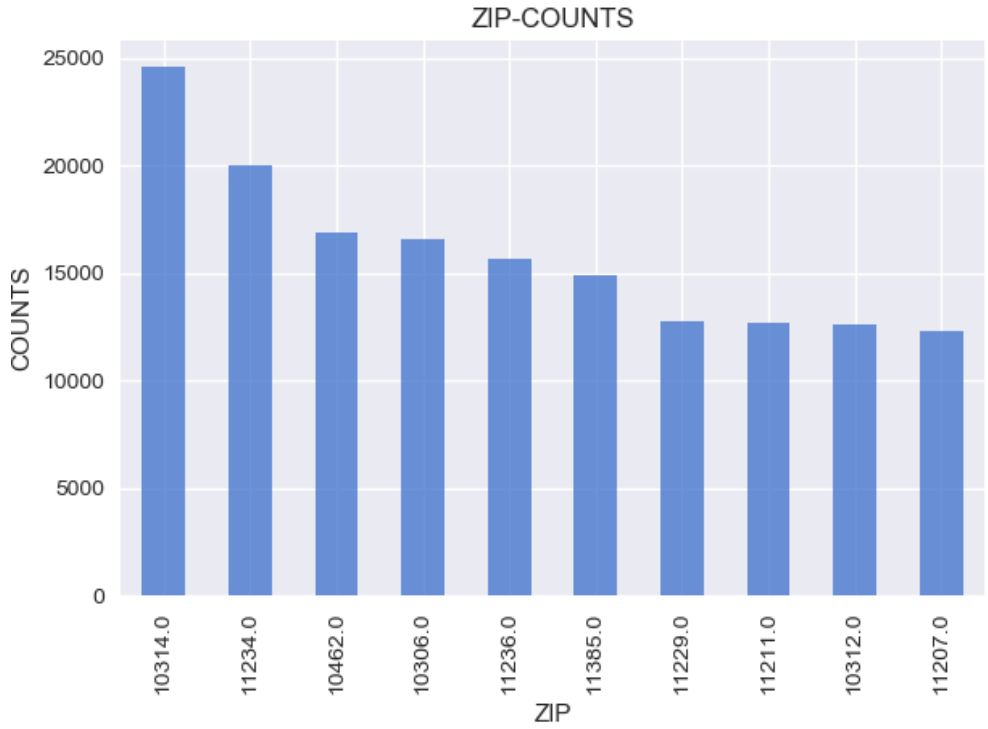
EXLAND (numeric variable): Value of the exempt land. There are 33186 unique values and no missing values.



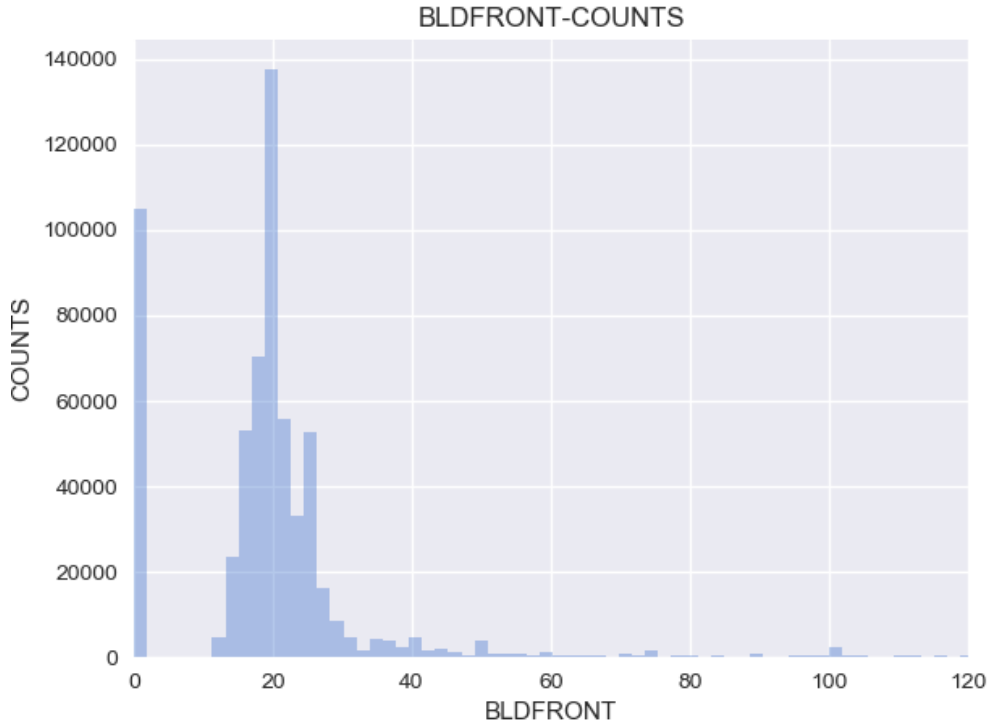
EXTOT (numerical variable): Value of the exempt property. There are 63805 unique values and no missing values.



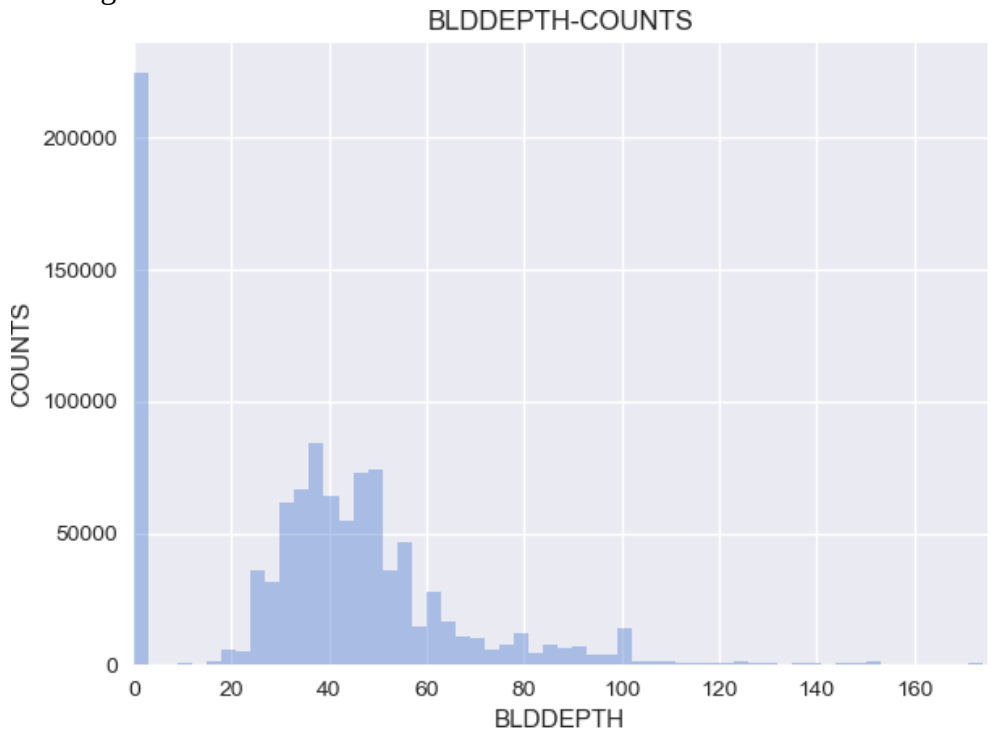
ZIP (categorical variable): Zip code of the property. There are 197 unique values and 26356 missing values.



BLDFRONT (numerical variable): The length of building frontage in feet. There are 610 unique values and no missing values.



BLDDEPTH (numerical variable): The length of building depth in feet. There are 620 unique values, no missing values.



❖ Part II: Data Preparation

2.1 Adjusting and combining existing variables

We created a new variable **BORO**, which is the first digit of **BBLE** to indicate the borough where the property located.

We defined a new variable **LTAREA** to illustrate total area of each property's lot by multiplying **LTFRONT** and **LTDEPTH**.

We defined a variable **BLDAREA** to illustrate total area of each property's building size by multiplying **BLDFRONT** and **BLDDEPTH**.

We also generated **BLDVOL** to show volume of each building by multiplying **BLDFRONT**, **BLDDEPTH** and **STORIES**.

Another new variable that we created is **MISSING_COUNTS**, which stands for the number of missing values of each row before we filled them up.

2.2 Removing Variables

We removed three types of variables: less informative variables, less populated variables, and variables already being aggregated.

Firstly, we found 8 less informative variables in total - **PERIOD, YEAR, VALTYPE, STADDR, OWNER, BLOCK, LOT, BLDGCL**. For **PERIOD, YEAR, VALTYPE,** all the observations shared the same value, which couldn't provide valuable information. In terms of **STADDR** and **OWNER**, they did present essential identification information of the records. However, there were 847055 unique values in **OWNER** and 820638 unique values in **STADDR**, which also included repetitive values hard to be identified. It would be hard and inefficient to feed these two variables into our model. In terms of **BLOCK** and **LOT**, they were not unique within each **BORO**. Thus they couldn't be regarded as informative identifier for each property. Although **BLDGCL** represented building class, we considered it as less informative variable, because the **BLDGCL** classification corresponded to tax class totally and we took tax class instead. After identifying the less informative variables, we removed all of them during the data cleaning process.

Next, we identified seven less populated variables - **EXCD1, EXMPTCL, AVLAND2, AVTOT2, EXLAND2, EXTOT2, EXCD2**. Since they are less populated and also cannot

indicate strong actual meaning for fraud identification, we removed them all. And the populated rates are as follow:

Variable	% of Populated
EXCD1	59.38%
EXMPTCL	1.43%
AVLAND2	26.80%
AVTOT2	26.80%
EXLAND2	8.27%
EXTOT2	12.39%
EXCD2	8.67%

Third, since we created three new variables **LTAREA, BLDAREA,** and **BLDVOLUME**. Taking these new aggregated variables in our following analysis, we decided to remove **LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH,** and **STORIES** from our dataset.

To sum up, we removed 15 variables in total. Meanwhile, we created and aggregated 3 new variables - **LTAREA, BLDAREA,** and **BLDVOLUME**.

2.3 Filling in the missing values and zero values

2.3.1 Numeric variables

To prepare for variable construction, it is necessary to replace zeros with meaningful numbers for the following variables, which are going to be divided by **FULLVAL, AVLAND, AVTOT** for the future analysis.

For the variable **LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES**, we replace zeros with the average **LTFRONT, LTDEPTH, BLDFRONT, BLDDEPTH, STORIES** by **BORO**.

2.3.2 Categorical variables

For the variable **ZIP**, there are 2.5% records with missing values. Thus we filled in the missing values with the average **ZIP** by **BORO** since **ZIP** is continue in each **BORO**.

For the variable **EASEMENT**, there are 99.6% records with missing values, which means they don't have an easement type in the record. Since the easement type is highly correlated with the value of properties, thus we filled in the missing values with the new category **'0'**.

❖ Part III: Variable Construction

Before generating new variables, we divided the original variables into two groups to analyze: numerators and denominators. After adjustment for existing variables, **13** numerators and **5** denominators are in hand:

The **13** numerators variables are:

1. **FULLVAL**: full value of the property
2. **AVLAND**: assessed value of land
3. **AVTOT**: assessed value of total property
4. **FULLVAL/LOTAREA**: the ratio of full value of property to the size of property's lot
5. **AVLAND/LOTAREA**: the ratio of assessed value of land to the size of property's lot
6. **AVTOT/LOTAREA**: the ratio of assessed value of property to the size of property's lot
7. **FULLVAL/BLDAREA**: the ratio of full value of property to the area of property's building
8. **AVLAND/BLDAREA**: the ratio of assessed value of land to the area of property's building
9. **AVTOT/BLDAREA**: the ratio of assessed value of property to the area of property's building
10. **FULLVAL/BLDVOL**: the ratio of property full value to the volume of property's building
11. **AVLAND/BLDVOL**: the ratio of land assessed value to the volume of property's building
12. **AVTOT/BLDVOL**: the ratio of property assessed value to the volume of property building
13. **MISSING_COUNTS**: the number of missing value of each property before being filled up

All numerators are numeric variables, which closely relate to the monetary value of properties. First three variables are from original dataset. The variable 4-13 are created through the procedure described in the 2.1.

The **5** denominator variables are:

1. **BORO**: borough code
2. **ZIP3**: the first three digits of zip code
3. **ZIP5**: the first five digits of zip code
4. **TAXCLASS**: tax class
5. **EASEMENT**: An easement is a nonpossessory right to use and/or enter onto the real property of another without possessing it.

All the denominator variables are used to classify numerators **except MISSING_COUNTS**. That is, we grouped all those numerators by these denominator variables, calculated mean of numerical variables in each group, and divided numerical variables by the mean of each group.

For example, if we combine **FULLVAL / LOTAREA** (numerator) and **TAXCLASS** (denominator), the expert variable would be (**FULLVAL / LOTAREA** of the record) / (mean of **FULLVAL / LOTAREA** in the **TAXCLASS** that the property belongs).

In total, we created **64** expert variables.

❖Part IV: Principal Components Analysis

After constructing 64 variables, we conducted Principal Components Analysis (PCA) for dimensionality reduction. PCA is an approach for deriving a low-dimensional set of features from a large set of variables, which is used to perform dimensionality reduction to avoid the curse of dimensionality.

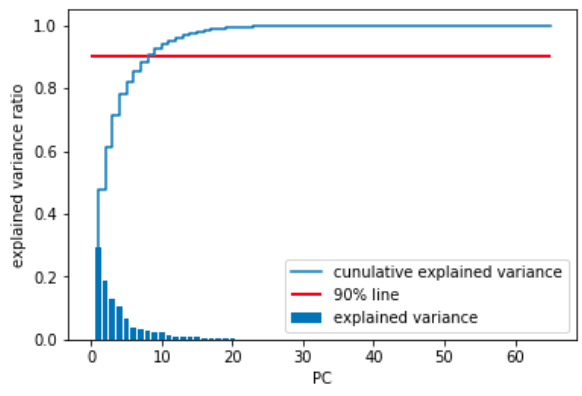
Prior to implementing PCA, standardization must be conducted so that all the variables were measured based on the same scale and those variables with large absolute values and big scales wouldn't overwhelm PCA. Therefore, we standardized variables first using **scale()** function in R. By default, **center = TRUE** set the mean at 0 of all the variables. Furthermore, by setting **scale = TRUE**, we could scaled the variables having standard deviation of 1. Using a for loop, we z-scaled all the variables of mean at 0 and standard deviation at 1.

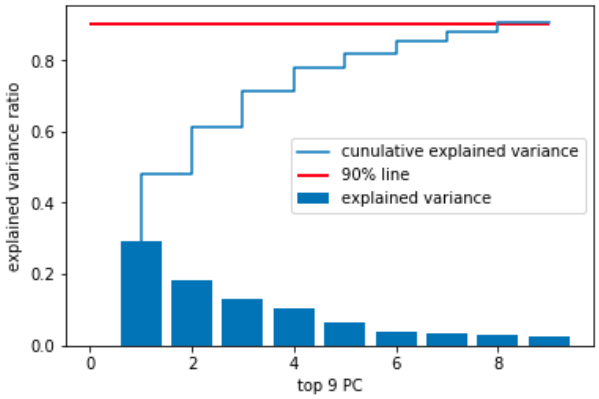
After the process of standardization, we use **PCA()** function (setting **graph=TRUE**) in the package **FactoMineR** to perform principal components analysis (PCA) and **eigen()** function to derive the eigenvalues and eigenvectors.

Following table shows the eigenvalue and cumulative variance explained.

	eigenvalue	cumulative percentage of variance
comp 1	19.1110229	0.29401574
comp 2	12.0422652	0.47928135
comp 3	8.53500583	0.61058914
comp 4	6.74526295	0.71436241
comp 5	4.26064362	0.77991078
comp 6	2.58459227	0.81967373
comp 7	2.20966126	0.85366852
comp 8	1.82454371	0.88173843
comp 9	1.65830725	0.90725085

The following two plots show the explained variance ratio of all PCs and explained variance ratio of top 9 PCs





By looking at the cumulative variance explained, we learn that the first PC explains 29.4% of variance, the top 9 PCs explain 90.7%. Thus, we decided to keep PC1 to PC9, which explained over 90% of the variance..

Finally, we standardized selected PCs for subsequent analysis.

❖ Part V: Fraud Algorithm

To quantify the extent of deviation, we chose Euclidean distance, Manhattan distance and autoencoder as our score algorithms. After scoring the observations, we binned each score into 1000 subgroups, and calculated final weighted score accordingly.

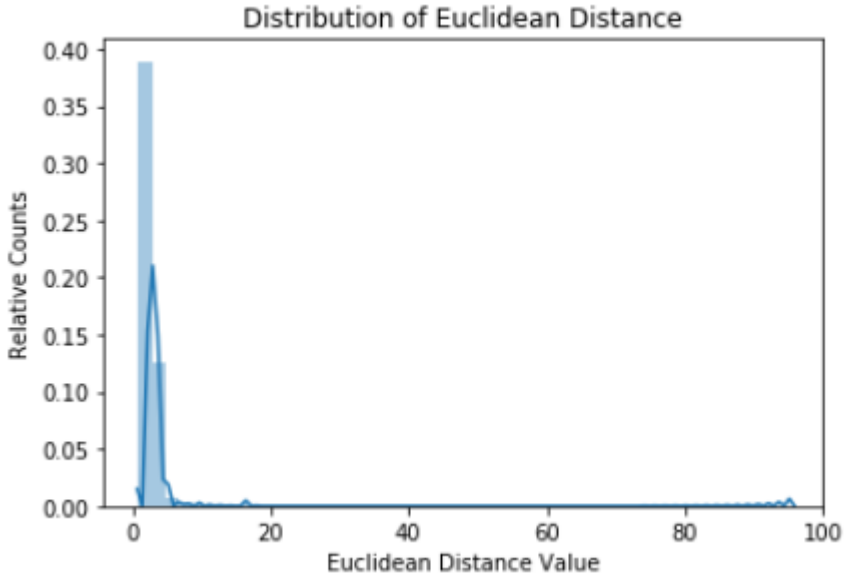
5.1 Distance

5.1.1 Euclidean Distance

In mathematics, the Euclidean distance or Euclidean metric is the "ordinary" straight-line distance between two points in Euclidean space. In this case, the measurement is:

$$EuclideanDistance = \sqrt{PC_0^2 + PC_1^2 + PC_2^2 + .. + PC_8^2}$$

The distribution of Euclidean scores is shown as follow, a clear right skewed, long tail pattern can be observed. This pattern matches our anticipation of fraud score distribution.

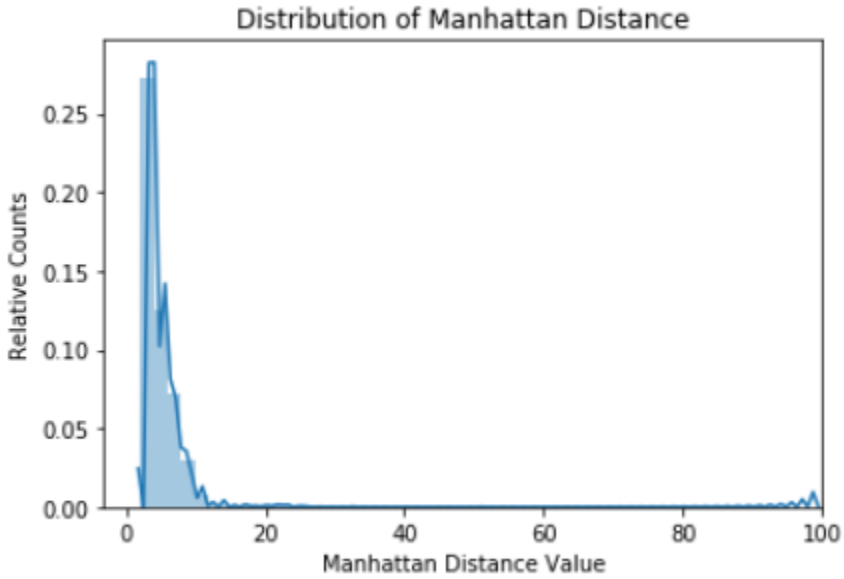


5.1.2 Manhattan Distance

The Manhattan distance is the simple sum of the absolute values of the differences of the coordinates. In this case, the measurement is:

$$ManhattanDistance = |PC_0| + |PC_1| + |PC_2| + .. + |PC_8|$$

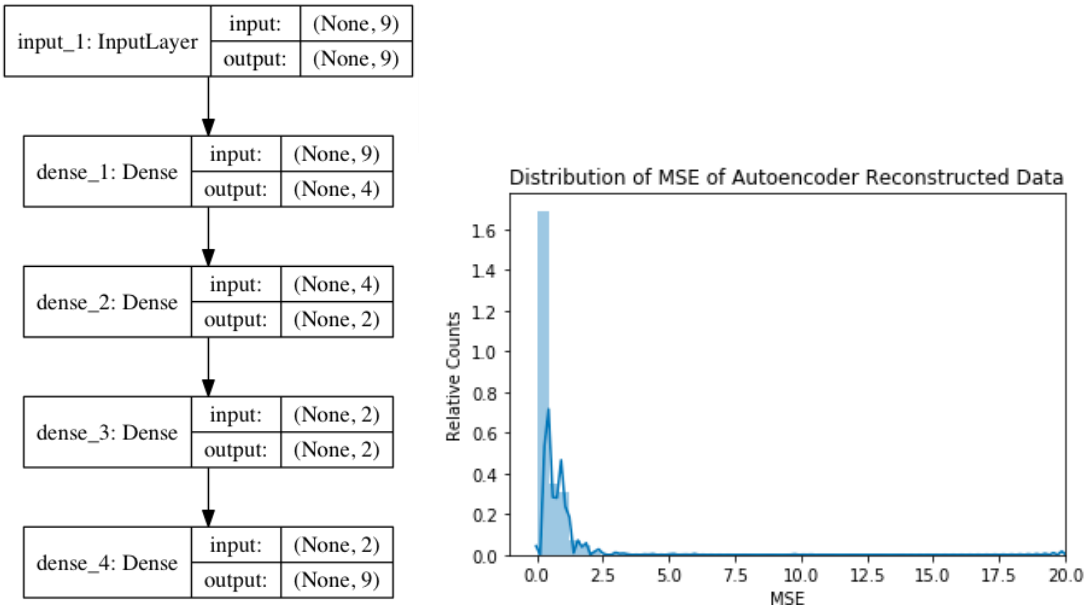
The distribution of Euclidean scores is shown as follow. Again, the right skewed, long tail pattern appears as expected.



5.2 Autoencoder

Autoencoder is an unsupervised learning technique that can be used to find the typical representation of inputs. To serve our purpose of finding potential frauds, we specifically use autoencoder for outlier detection. In other words, if a type of inputs is densely existing in the dataset, we can expect it to be reconstructed well after encoding and decoding; If a type of inputs is rare in the dataset, we can expect it to be reconstructed poorly. Our selected measurement of reconstruction quality is MSE (mean squared error) between reconstructed value and real value.

As expected, our MSE score showed right skewed long tail pattern as the distance algorithms above.



Left: structure of autoencoder; Right: MSE distribution of autoencoder

5.3 Weighted Score

Based on the distribution of scores, we decided to set our weighted score a linear combination of the three binned scores mentioned above.

Weighted Score = 0.3 * Manhattan Distance + 0.4 * Euclidean Distance +
0.3 * Autoencoder MSE

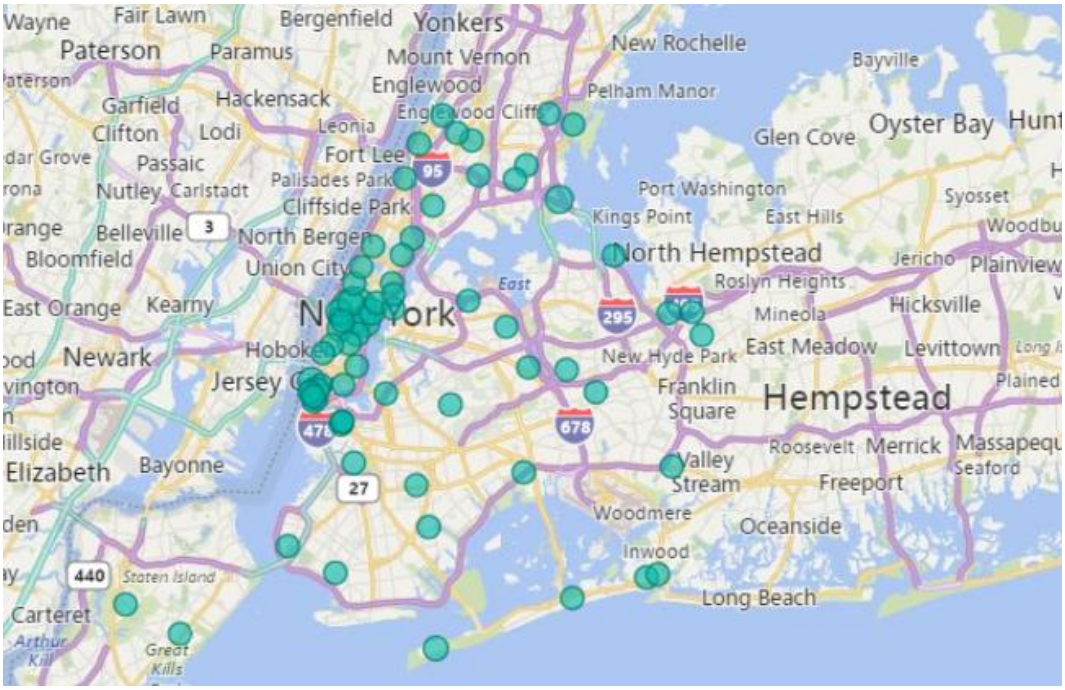
By filtering **Weighted Score = 1000**, we extracted 669 outlier records for further investigation.

❖ Part VI: Results

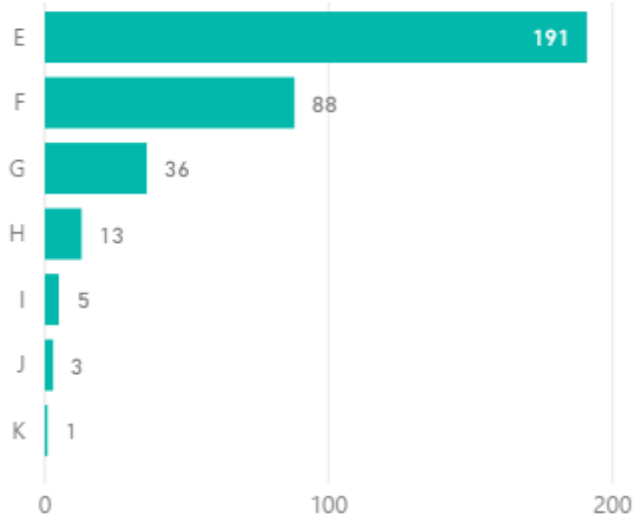
6.1 Descriptive Conclusion

Some characteristics of the 669 suspicious records extracted by weighted score above are shown below:

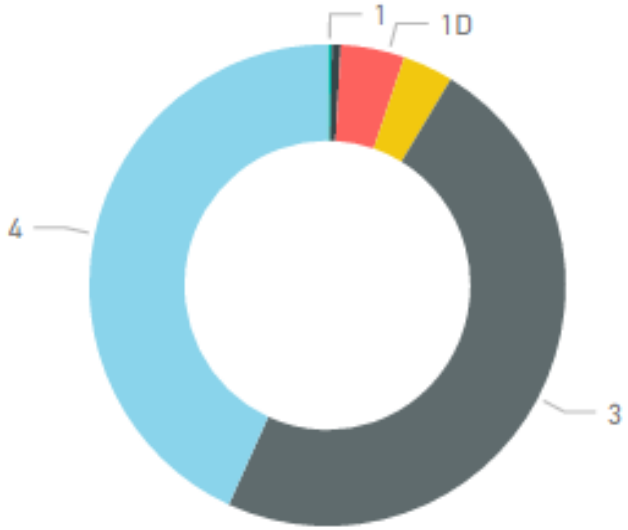
- The location of most of the suspicious observations are waterfront, as can be shown in the following map



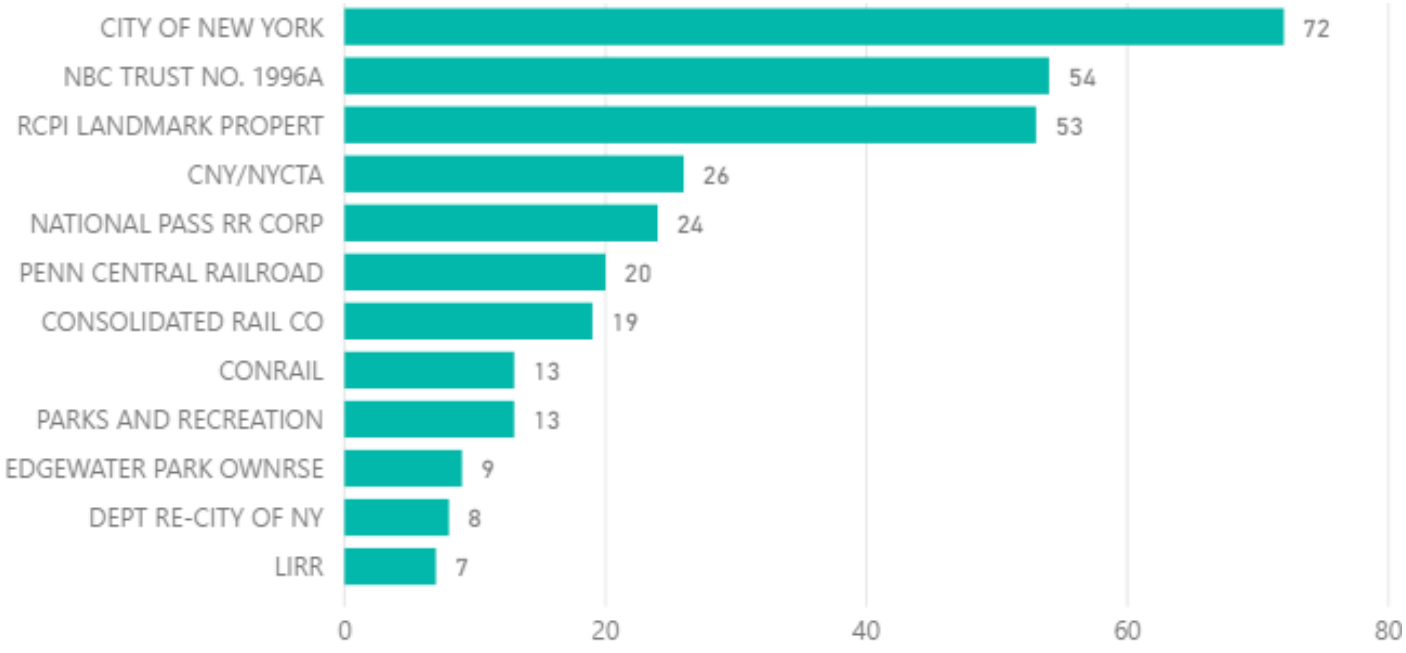
- Top two **EASEMENT** types concerning occurrence for the suspicious observations are ‘E’ and ‘F’.



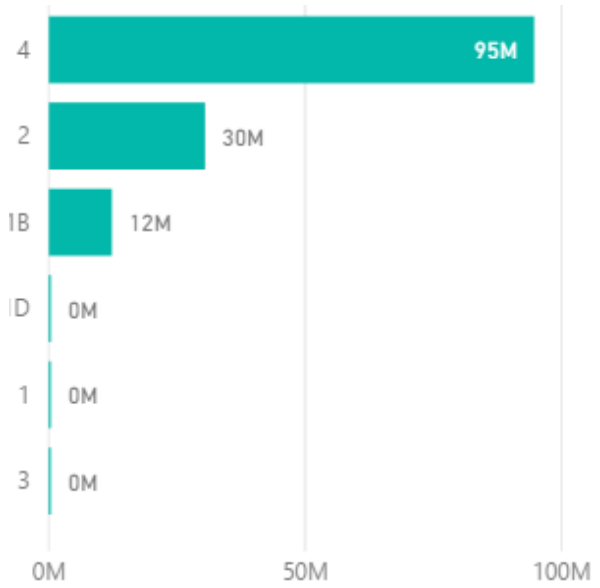
- For **TAXCLASS**, 3 and 4 made up more than 90% of the suspicious records.



- For the **OWNER**, City of New York, NBC Trust NO. 1996A and RCPI Landmark Property have 179 records altogether. Moreover, all 54 records for NBC and 53 records for Landmark are located at zip code 10020.



- If we look at the average of **EXTOT** by **TAXCLASS**, we can find that although **TAXCLASS 3** has the highest number of records, **TAXCLASS** '4' has the highest average value of **EXTOT**.



- **TAXCLASS** for top 10 **OWNER** concerning average of **EXTOT** are all 4.

OWNER	Average of EXTOT	TAXCLASS
LOGAN PROPERTY, INC.	4,668,308,947.00	4
CULTURAL AFFAIRS	1,537,650,000.00	4
U S GOVERNMENT OWNRD	1,347,458,332.50	4
NEW YORK STATE DEPART	1,318,275,000.00	4
NY CONVENTION CTR DVL	515,250,000.00	4
ONE BRYANT PARK	504,000,000.00	4
DEPT OF GENERAL SERVI	445,736,130.00	4
THE PORT AUTHORITY OF	292,500,000.00	4
CITY OF NEW YORK	268,117,800.00	4
UDC	251,775,000.00	4
DORMITORY AUTHORITY O	224,158,500.00	4
NYS URBAN DEVELOPMENT	212,850,000.00	2
BRONX V A MEDICAL CEN	210,645,000.00	4
TIMES SQUARE HMC HOTE	199,300,000.00	4
MOUNT SINAI HOSPITAL	193,950,000.00	4

6.2 Exempt Properties Without Owner Name

A common property fraud is exemption fraud. According to our output, the **STADDR** that contains the most outliers is **30 ROCKEFELLER PLAZA**, which is part of New York City government involved landmark project.

With this connection of governmental involvement and tax exemption in mind, we tried to filter out those properties with **EXTOT**(tax exemption) but without **OWNER** registered as below. These 11 outliers deserves further investigation.

RECORD	BBLE	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	ZIP
103668	1000160230	16	230	NA	NA	D3	2	196	10282
228666	1000160235	16	235	NA	NA	D1	2	260	10282
241149	1012681102	1268	1102	NA	NA	R5	4	200	10019
458276	1000160190	16	190	NA	NA	D8	2	196	10282
559023	2024930001	2493	1	NA	NA	Q6	4	798	10451
581492	1000161301	16	1301	NA	NA	R5	4	0	10282
694876	4017870020	1787	20	NA	NA	Q6	4	1700	11368
805688	1017750003F	1775	3	F	NA	V9	4	60	NA
902909	1000160185	16	185	NA	NA	D8	2	196	10282
907040	1000160195	16	195	NA	NA	D8	2	196	10282
1037511	1000161302	16	1302	NA	NA	R5	4	0	10282

RECORD	LTDEPTH	STORIES	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT	EXCD1	STADDR
103668	101	22	22941000	900000	10323450	900000	900000	6800	300 NORTH END AVENUE
228666	179	32	17845000	1620000	8030250	1620000	1620000	6800	200 NORTH END AVENUE
241149	315	40	483000000	80734554	217350000	0	1722900	1985	666 5 AVENUE
458276	71	NA	73100000	1921500	32895000	1921500	32895000	6800	22 RIVER TERRACE
559023	611	6	1663775000	78750000	748698750	78750000	748698750	2500	1 EAST 161 STREET
581492	0	14	99580000	5363820	44811000	5363820	44811000	6800	102 NORTH END AVENUE
694876	1635	6	1223500000	110025000	550575000	110025000	550575000	2500	123-01 ROOSEVELT AVENUE
805688	115	NA	9010	4055	4055	4055	4055	2201	NA
902909	70	24	70800000	2601000	31860000	2601000	31860000	2191	211 NORTH END AVENUE
907040	135	24	73800000	4500000	33210000	4500000	33210000	6800	325 NORTH END AVENUE
1037511	0	14	37500000	2700000	16875000	2700000	16875000	6800	102 NORTH END AVENUE

RECORD	EXMPTCL	BLDFRONT	BLDDEPTH	AVLAND2	AVTOT2	EXLAND2	EXTOT2	EXCD2
103668	NA	0	0	747204	10170654	747204	747204	NA
228666	NA	0	0	1377081	7787331	1377081	1377081	NA
241149	NA	200	315	80734554	226383703	NA	1722900	NA
458276	NA	0	0	1921500	32868000	1921500	32868000	NA
559023	NA	0	0	NA	NA	NA	NA	NA
581492	NA	0	0	5363820	43583200	5363820	43583200	NA
694876	NA	800	170	95245199	547348274	95245199	547348274	NA
805688	NA	0	0	2640	2640	2640	2640	NA
902909	NA	196	70	2511450	28799550	2511450	28799550	5116
907040	NA	0	0	4500000	38421000	4500000	38421000	NA
1037511	NA	0	0	2700000	15717000	2700000	15717000	NA

6.3 Top-10 Scored Records

Based on our three algorithms, top 10 scored records of Euclidean method were exactly the same as top-10 of Autoencoder method. And 9 of them were overlapped with top 10 records of

Manhattan method. In the following analysis, we would examine and provide analysis for each of top 10 records based on the data.

The table below shows top 10 records based on our original dataset:

RECORD	BBLE	BLOCK	LOT	EASEMENT	OWNER	BLDGCL	TAXCLASS	LTFRONT	LTDEPTH
294060	4014340023	1434	23	NA	ACUNA, ROMEO E	A5	1	30	100
376242	4047280076	4728	76	NA	RACANELLI, JOSEPHINE	B2	1	19	119
78803	3071420005	7142	5	NA	SIDOTI FRANK B	B9	1	18	100
22920	4039161001	3916	1001	NA	ERIK MICHAEL SANCHEZ	R3	1A	16	85
315452	1017750006K	1775	6	K	HOUSING & DEV ADMIN	V1	4	0	0
180683	4126830023	12683	23	NA	SNAPE, GENE R	A2	1	31	100
270464	1001421598	142	1598	NA	UNIT 1040 ASSOCIATES,	R4	2	0	0
651202	4048610043	4861	43	NA	VICARI, GIAN PIERO	A2	1	42	95
616392	3011730074	1173	74	NA	M & M CENTURY GROUP L	C1	2	50	123
447395	5016700323	1670	323	NA	JOSEPH TESTAVERDI	A9	1	24	122

RECORD	STORIES	FULLVAL	AVLAND	AVTOT	EXLAND	EXTOT	EXCD1	AVTOT2	ZIP
294060	2.7	748000	20250	33735	0	0	NA	NA	11372
376242	2	743000	23620	39087	1620	1620	1017	NA	11357
78803	3	534000	14960	23155	1620	1620	1017	NA	11223
22920	1	261461	3425	13671	0	0	NA	NA	11356
315452	NA	22200	9990	9990	9990	9990	2201	4328	10035
180683	1.5	353000	7377	14630	0	0	NA	NA	11413
270464	32	337079	18444	151686	1502	134744	5110	148503	10007
651202	1	508000	19398	30135	3240	3240	1017	NA	11357
616392	4	484000	55800	217800	0	0	NA	236340	11238
447395	2	276400	10992	16423	1620	1620	1017	NA	10303

RECORD	EXMPTCL	BLDFRONT	BLDDEPTH	AVLAND2	STADDR	EXLAND2	EXTOT2	EXCD2	
294060	NA	20	35	NA	33-36 87 STREET	NA	NA	NA	
376242	NA	22	56	NA	155-03 16 DRIVE	NA	NA	NA	
78803	NA	18	38	NA	124 AVENUE V	NA	NA	NA	
22920	NA	0	0	NA	121-39 5 AVENUE	NA	NA	NA	
315452	NA	0	0	4328	107 EAST 126 STREET	4328	4328	NA	
180683	NA	18	36	NA	120-42 198 STREET	NA	NA	NA	
270464	NA	0	0	18444	101 WARREN STREET	1555	131614	NA	
651202	NA	26	42	NA	22-04 160 STREET	NA	NA	NA	
616392	NA	50	90	45036	391 ST JOHN'S PLACE	NA	NA	NA	
447395	NA	16	51	NA	28 AMADOR STREET	NA	NA	NA	

(1) Record No.294060: The record has unusual values of **FULLVAL/ LTAREA** and **AVTOT/ LTAREA** with respect to **BORO, TAXCLASS, ZIP,** and **EASEMENT.**

(2) Record No.376242: The record has unusual values of **FULLVAL/ LTAREA** and **AVLAND/ LTAREA** with respect to **BORO, TAXCLASS, ZIP,** and **EASEMENT.** The ratios were significantly high, which may indicate the fraud.

(3) Record No.78803: The record has unusual values of **FULLVAL/ LTAREA** and **FULLVAL/ BLDAREA** with respect to **ZIP5**. The ratios were significantly high, which may indicate the fraud. We could also notice that it belonged to **TAXCLASS** 1 and got 1620 exemption.

(4) Record No. 22920: The record has unusual values of **FULLVAL** and **AVLAND** with respect with **ZIP5**. Furthermore, it's well worth mentioning that the **AVLAND/LTAREA** with respect to **TAXCLASS** is extremely low, which may indicate fraud.

(5) Record No. 315452: This record has no **LTFRONT**, **LTDEPTH, STORIES, EXMPCLS, BLDFRONT** and **BLDDEPTH** data. Besides,**AVLAND2,AVTOT2,EXLAND2,EXTOT2** are of the same value, which indicate that there might be some fraud.

(6) Record No. 180683: This record has no **EXLAND** and **EXTOT** data**.** Besides, This record has unusual values of **FULLVAL/ LTAREA** and **AVLAND/ LTAREA** with respect to **BORO, TAXCLASS** and **EASEMENT.**

(7) Record No.270464: This record has no **LTFRONT**, **LTDEPTH**, **BLDFRONT** and **BLDDEPTH** data. But the strange point is that the total exemption of the property is high in both **EXTOT1** and **EXTOT2**, which indicate that there might be some fraud in inside.

(8) Record No.651202: Although the record has no missing values inside, the **FULLVAL** respect to **LOTAREA**, **BLDAREA** and **BLDVOL** is unusual. Also, the number of stories is only 1 but the area is super large but the **TAXCLASS** is 1 which means 1-3 unit residences. The lot area, building area and the function of the property does not match perfectly, which may be some kind of fraud.

(9) Record No.616392: In terms of average level of its **TAXCLASS, AVTOT/LTAREA** is relatively too high, but the **FULLVAL/VOLUME** and **FULLVAL/LTAREA** is too low**.** Therefore, it shows fraud.

(10) Record No. 447395: **FULLVAL** and **AVLAND** compare to the average level of various categories, but **AVTOT** is significantly higher than means.

❖ Conclusion

In order to analyze potential fraud records for more than 1 million New York City Properties, we followed a general process of analysis including data cleaning, variable construction, zero-mean normalization and dimensionality reduction, fraud algorithms' application, score calculation and potential fraudulent records identification. Python, R and Microsoft Power BI are used during the process as effective tools.

High scored records are analyzed emphatically since they are highly suspicious fraudulent properties. We provided a detailed description analysis and also laid our special stress on analyzing the particularity of the top ten records, which have high scores in both of the algorithms, according to their owners, address, and some other related factors. Exemption analysis is also given in order to figure out more details about the potential fraudulent records, and the logic about exemption frauds is also shown at length.

However, we still have many things that can be done in the future for this project. For example, we can fill up the missing values more accurately according to the properties' own addresses, building volumes, and lot area data. We can also add some more variables about the relationship among **AVLAND**, **AVTOT**, and **FULLVAL** to figure out the differences between the properties' estimated values and their actual values. Since **TAXCLASS** is a very important index according to assessed values, some frauds might happen to chase after monetary benefits by undervaluing the properties. In addition, since the time of the dataset is 2011 and the data source is a little bit old for us now. We can look for some new data now of the properties in New York City and compare them with the old data and furtherly analyze the changes during the 7 years in New York City.

❖ Appendix

Data Quality Report

1. Summary Statistics for Numerical Variables

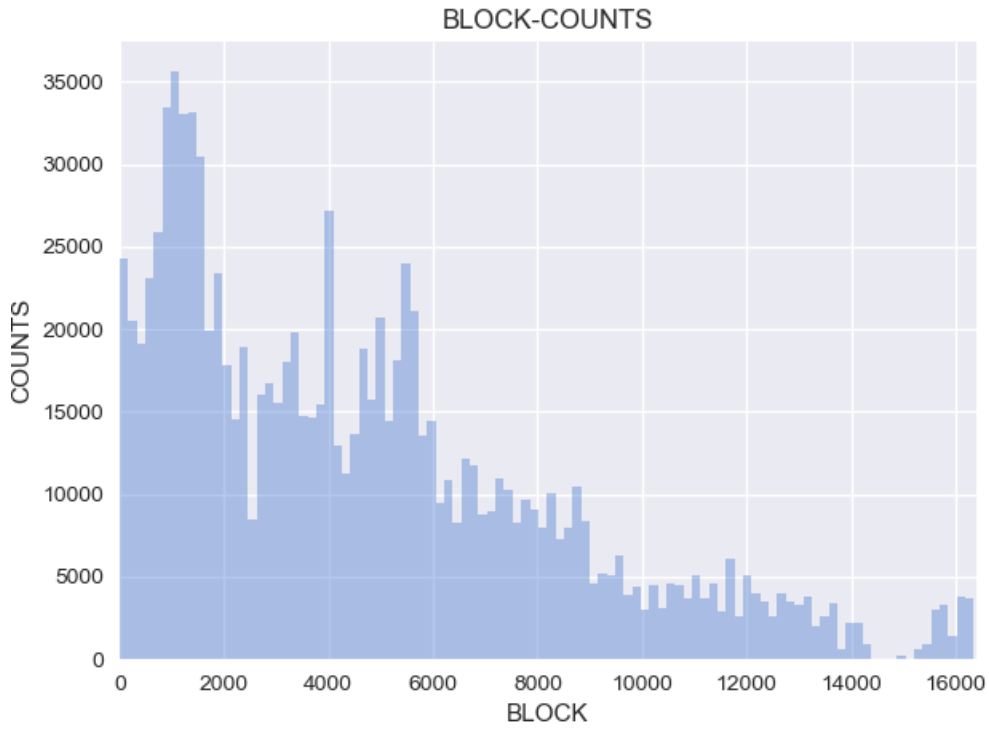
The numerical variables in the following chart are from NYC Open Data website (https://data.cityofnewyork.us/Housing-Development/Property-Valuation-and-Assessment-Data/rgy2-tti8). The mathematical statistics are provided, including minimum, mean, 25%, medium, 75% and maximum values, range, standard deviation, number of unique value and the percentage of population for each variable. The time of records is 2011.

Variable	Min	25%	Medium	75%	Max	Mean	Range	SD	Unique	%Populated
LTFRONT	0	19	25	40	9999	36.17	9999	73.72	1277	100
LTDEPTH	0	80	100	10	9999	88.27	9999	75.48	1336	100
STORIES	1	2	2	3	119	5.06	118	8.43	112	95.027
FULLVAL	0	303000	446000	619000	615000	880488	615000	11702930	108277	100
AVLAND	0	916000	136460	197060	266850	859950	266850	4100755	70529	100
AVTOT	0	183850	253390	460950	466850	230758	466850	6951206	112294	100
EXLAND	0	0	162000	162000	266850	368118	266850	4024330	33186	100
EXTOT	0	0	1620	2090	4.67E+09	92544	4.67E+09	6578281	63805	100
EXCD1	1010	1017	1017	1017	7170	1604.5	6160	1388.13	130	59.38
BLDFRONT	0	15	20	24	7575	23.02	7575	35.79	610	100
BLDDEPTH	0	26	39	51	9393	40	9393	43.04	620	100
AVLAND2	3	5705	20059	62338.75	2.37E+09	246366	2.37E+09	6199390	58170	26.795
AVTOT2	3	34014	80010	240792	4.50E+09	716079	4.50E+09	11690170	110891	26.796
EXLAND2	1	2090	3053	31419	2.37E+09	351802	2.37E+09	10852480	21997	8.266
EXTOT2	7	2889	37116	106629	4.50E+09	658115	4.50E+09	16129810	48107	12.391
EXCD2	1011	1017	1017	1017	7160	1371.66	6149	1105.49	61	8.673

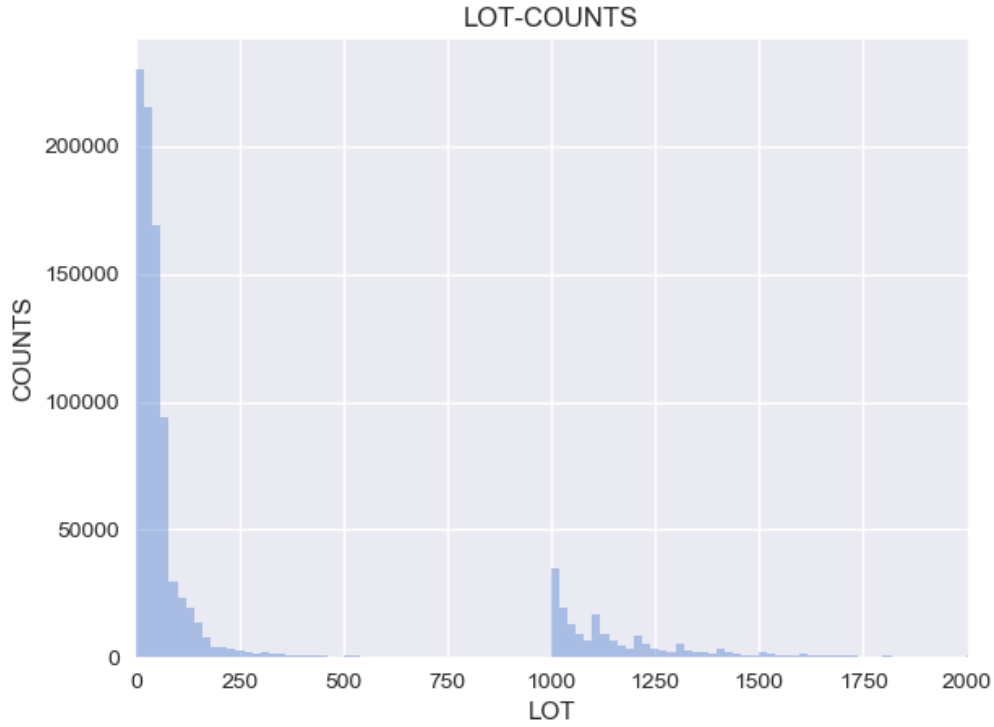
2. Detailed information for Each Field

Field Name	*Description*
RECORD	The number of each record. Discrete data with metric.

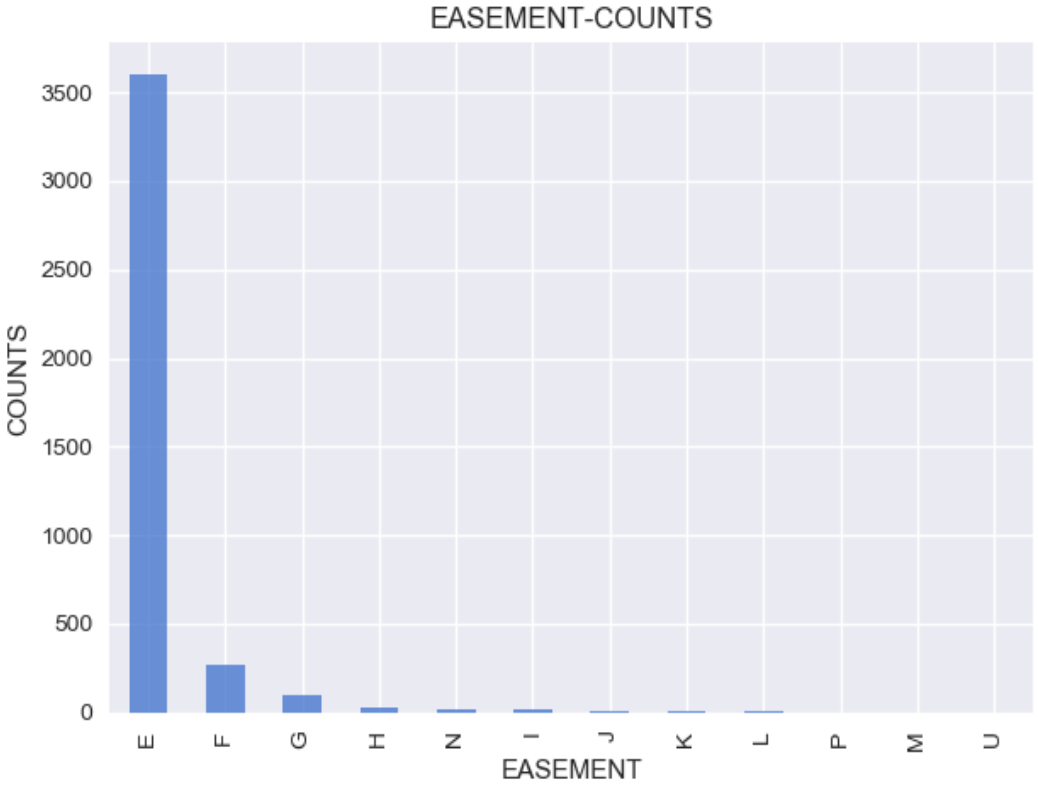
Field Name	*Description*
BLOCK	The area bounded by four streets. Continuous data with metric.

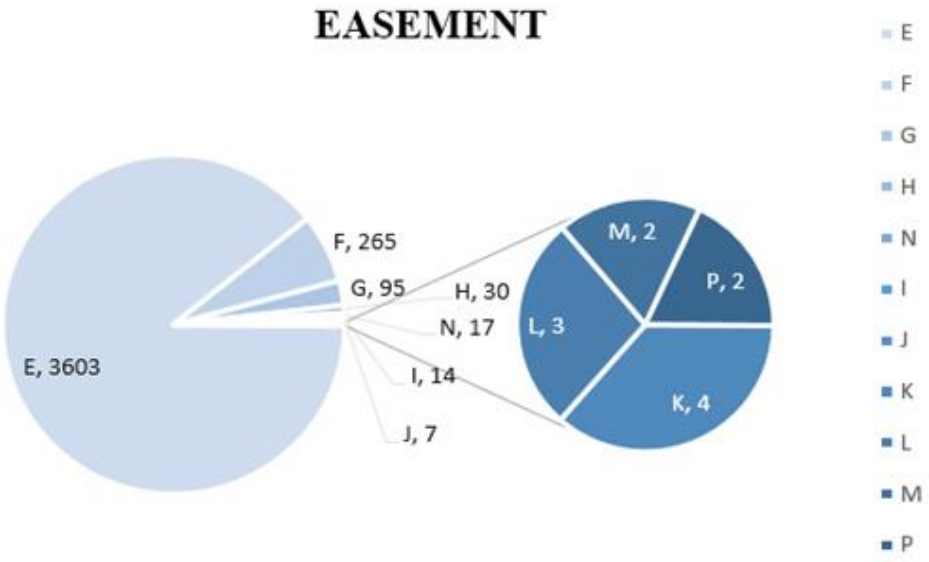


Field Name	*Description*
LOT	A plot of land assigned for sale or for a particular use. Continuous with metric.

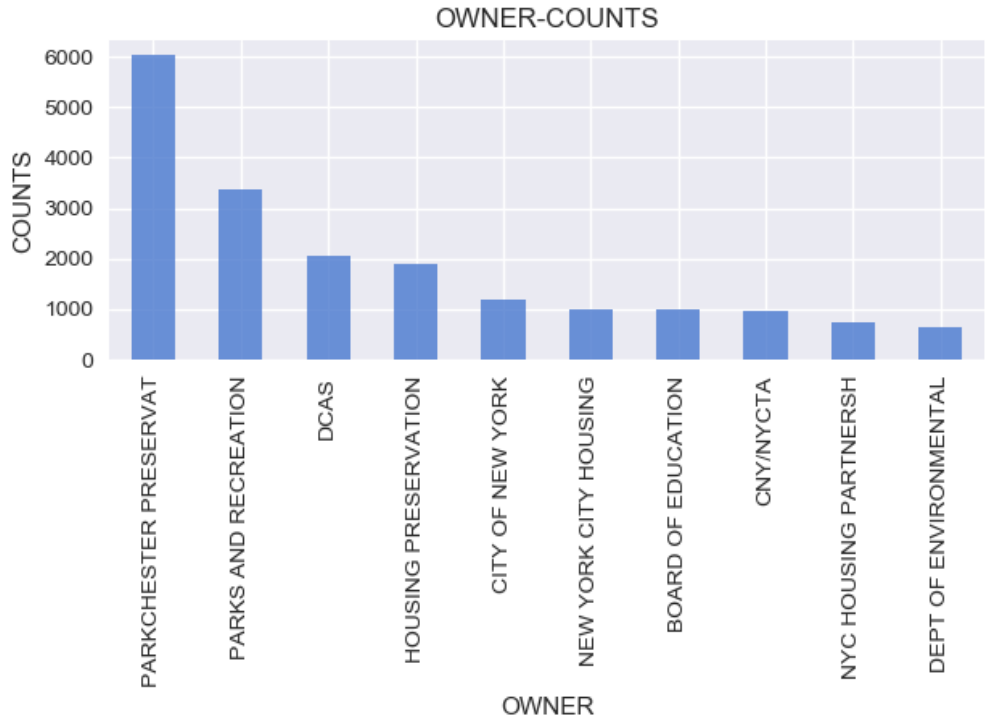


Field Name	*Description*
EASEMENT	A right to cross or otherwise use someone else's land for a specified purpose. Categorical data without metric.

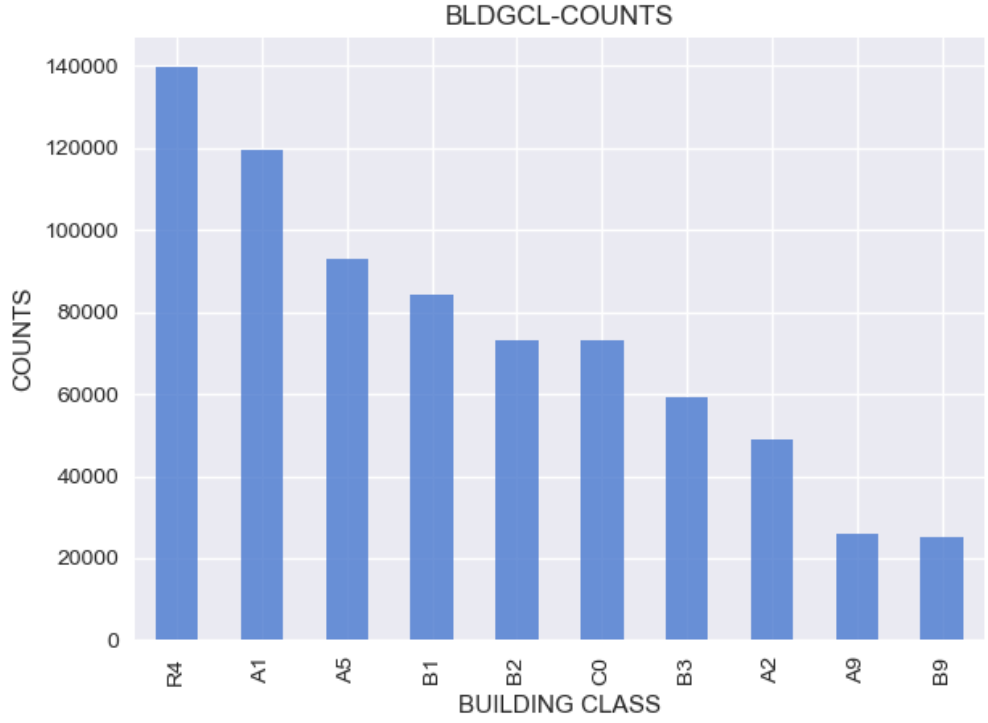




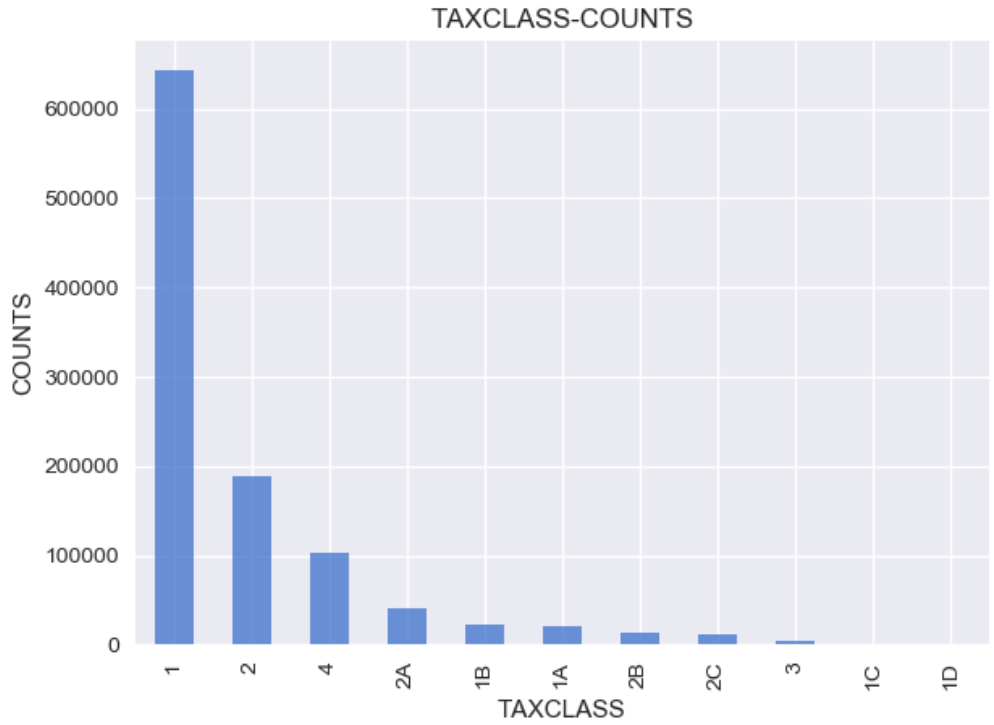
Field Name	*Description*
OWNER	An organization or a person who own the property. Characters without metric.



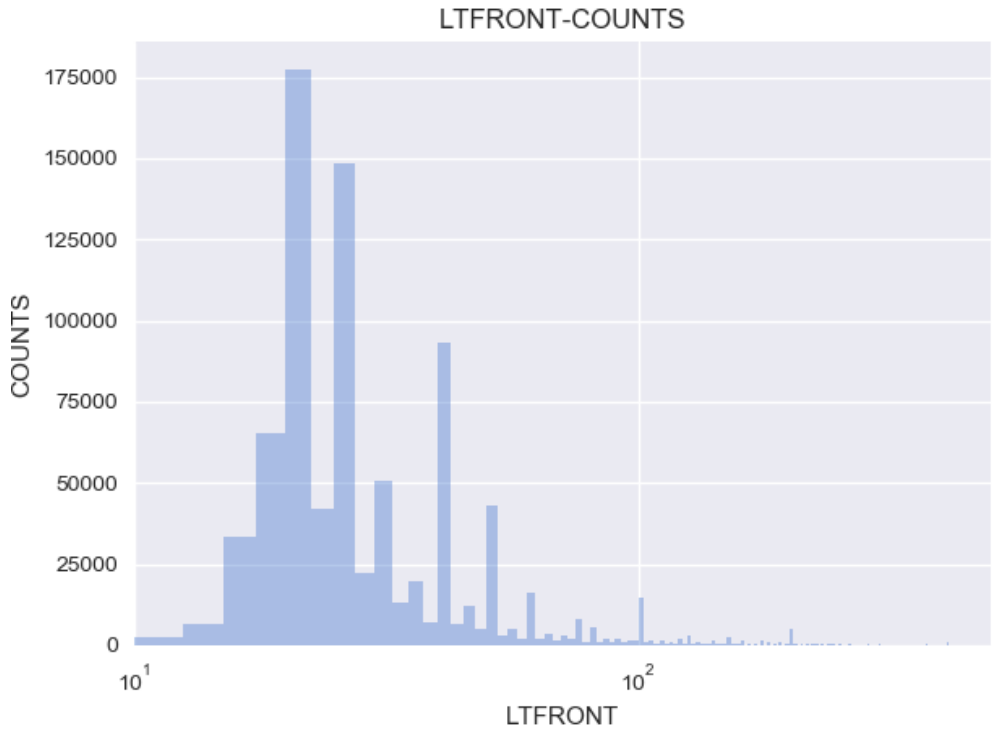
Field Name	*Description*
BLDGCL	Building class. Different building has different building classes and they are related to the tax classes directly. Categorical data without metric.



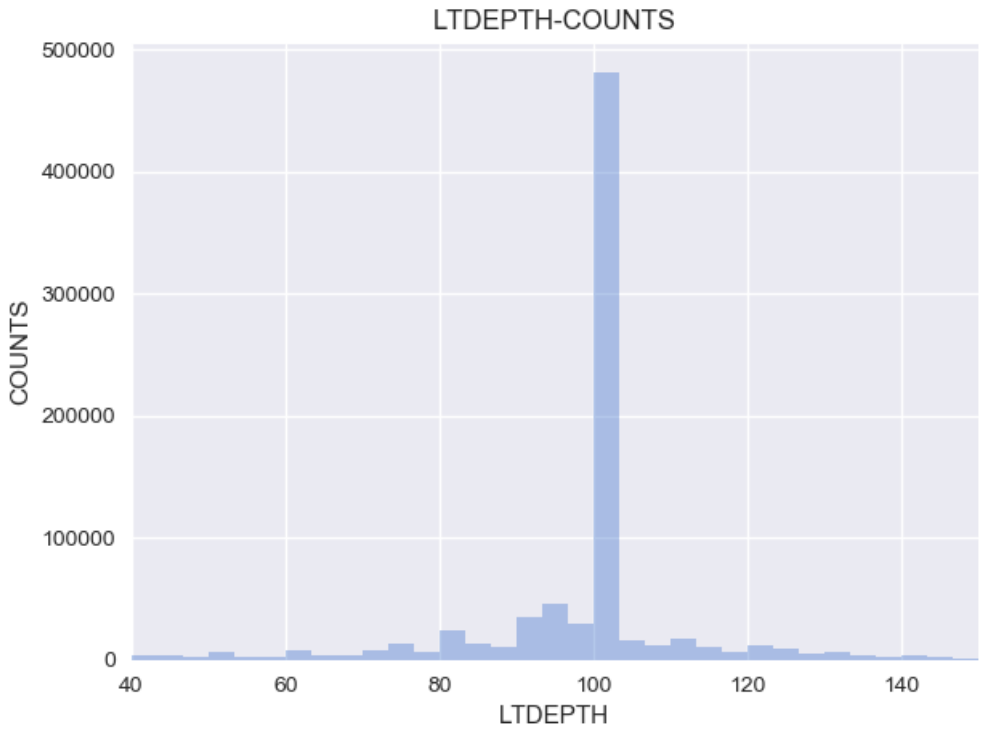
Field Name	*Description*
TAXCLASS	Tax class. It is used to define tax rules, which are combined by tax class, product tax class and tax rates. Categorical data without metric.



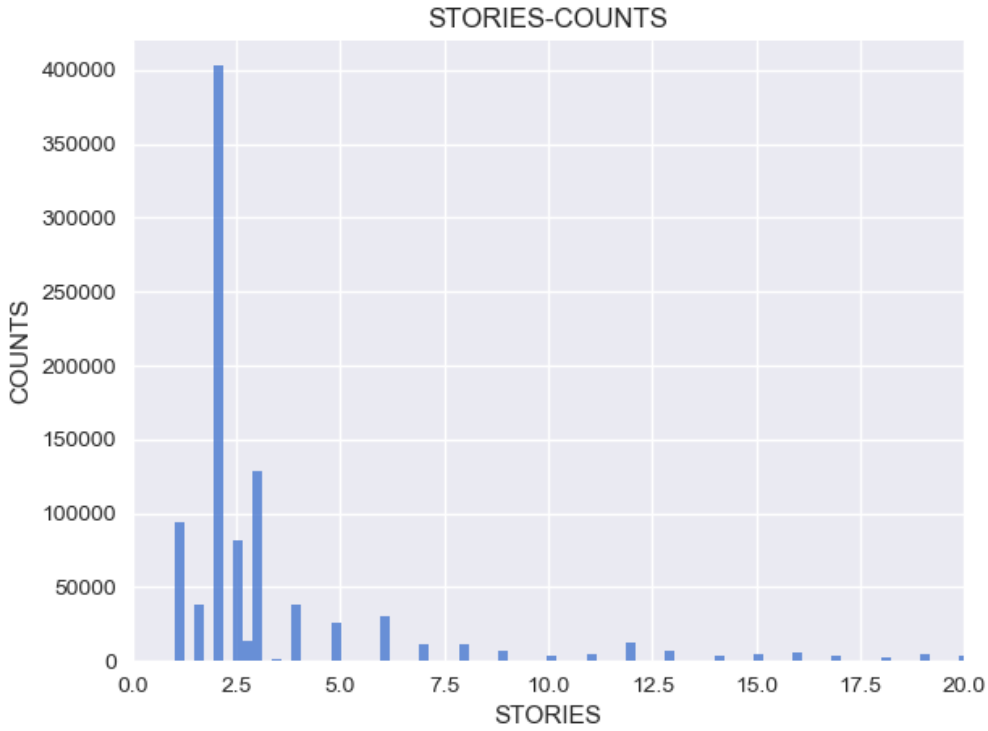
Field Name	*Description*
LTFRONT	Lot frontage, which means the length of the front lot line. Continuous with metric.



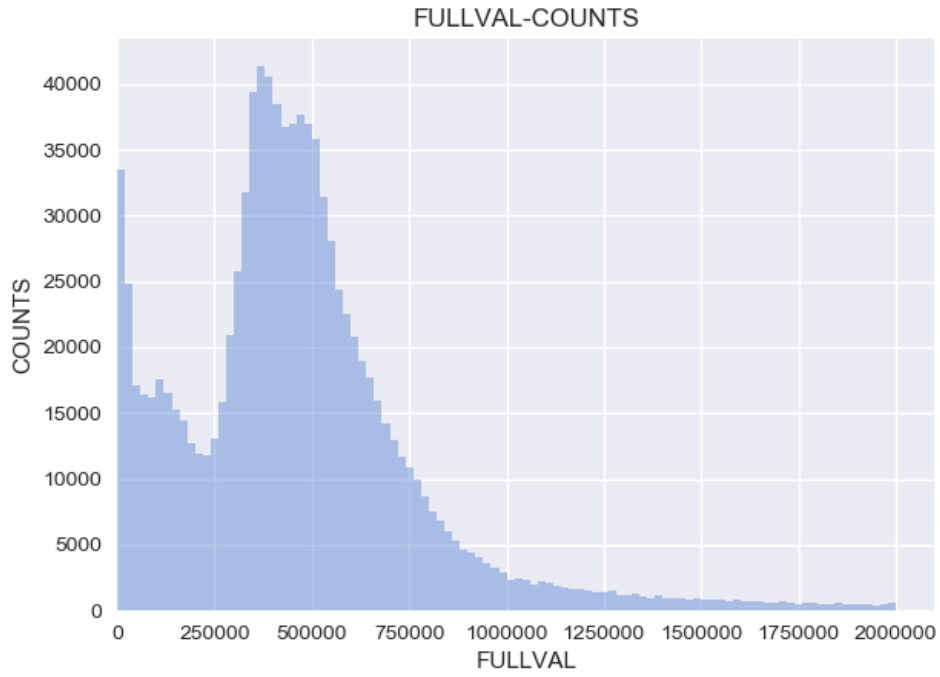
Field Name	*Description*
LTDEPTH	Lot depth means the horizontal distance between the front and rear property lines of a lot, measured along a line midway between the side property lines. Continuous with metric.



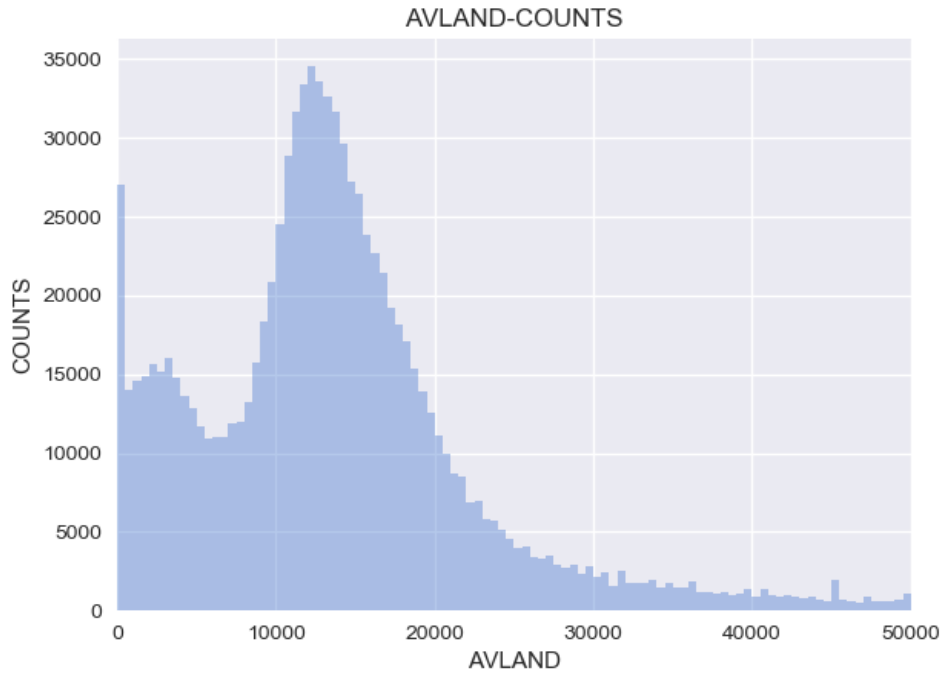
Field Name	*Description*
STORIES	A part of a building comprising all the rooms that are on the same level. Continuous data with metric.



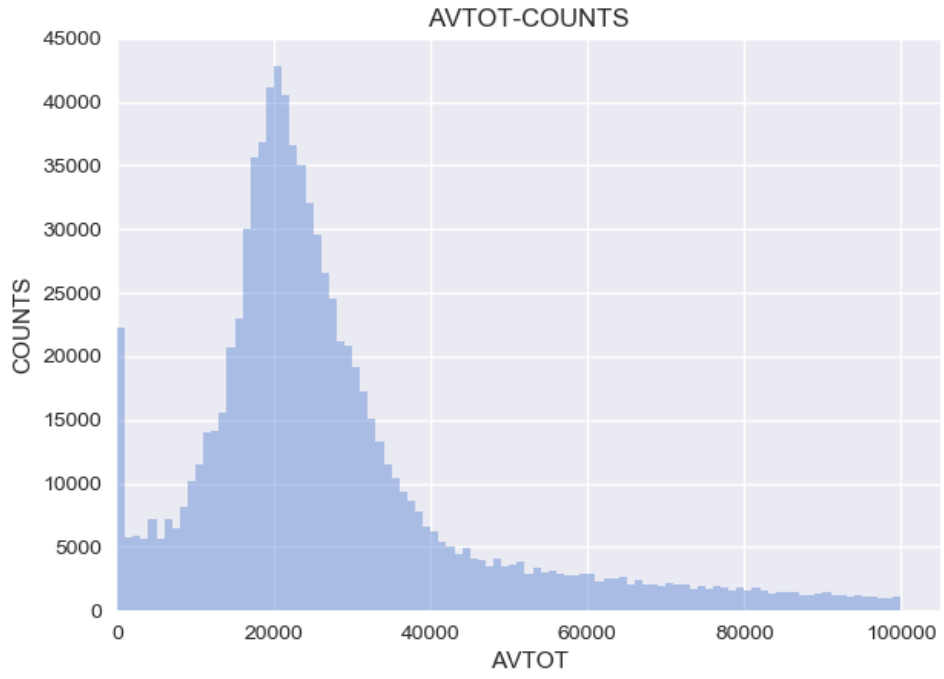
Field Name	*Description*
FULLVAL	The full values of the properties have in the market. Continuous data with metric.



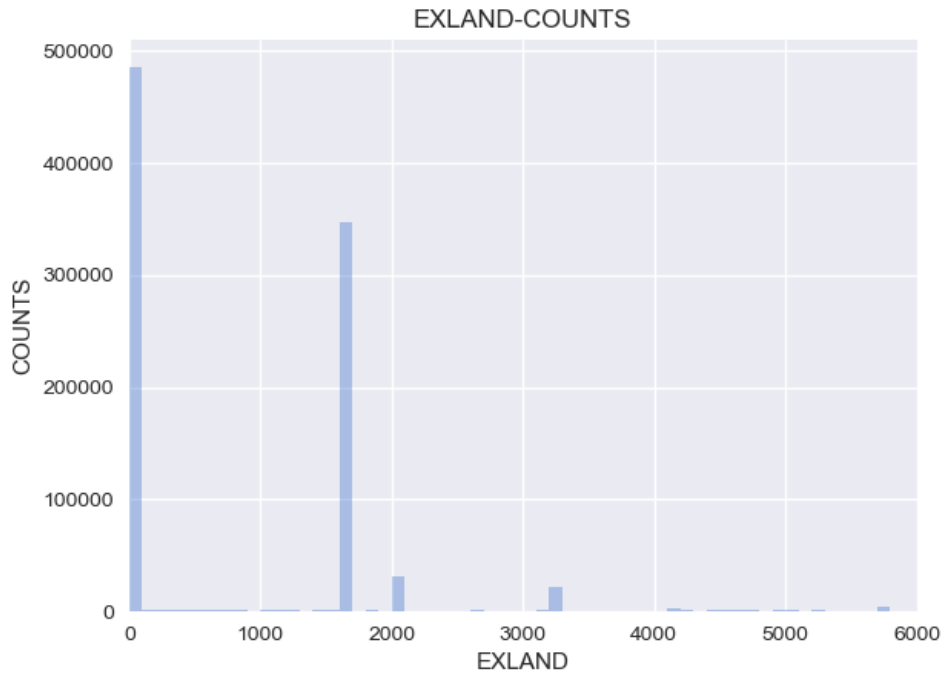
Field Name	*Description*
AVLAND	The assessed value of the properties' lands. Continuous data with metric.



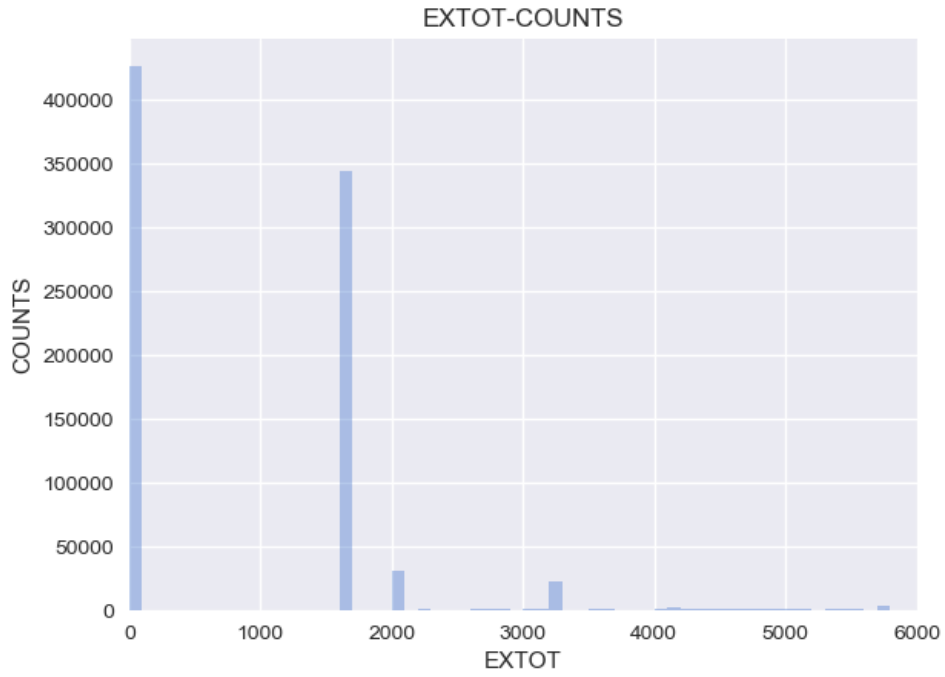
Field Name	*Description*
AVTOT	The assessed value of the total properties. Continuous data with metric.



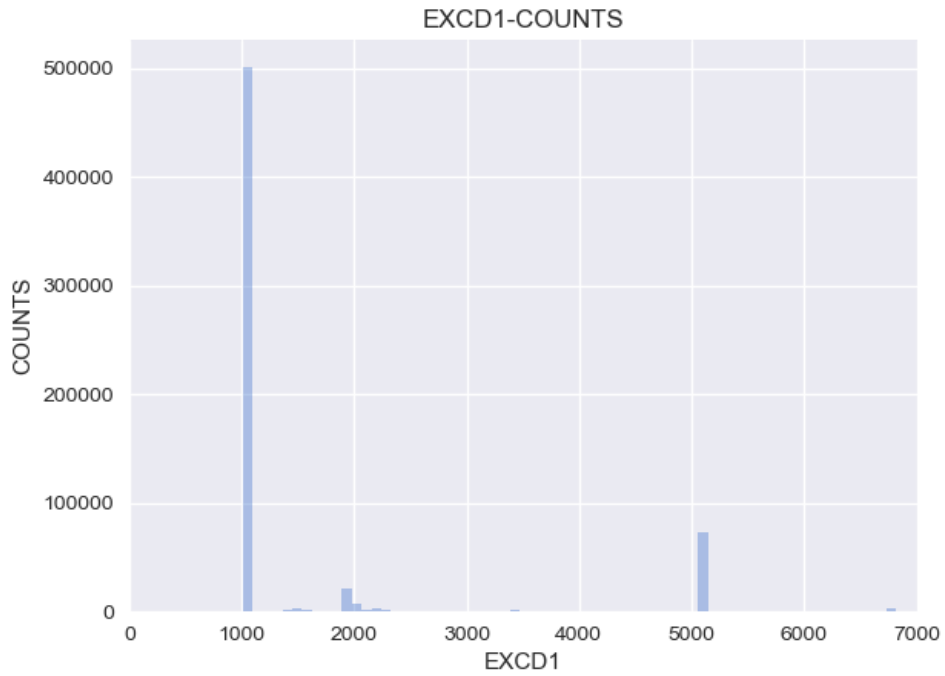
Field Name	*Description*
EXLAND	The exempt value of the properties' lands. Continuous data with metric.



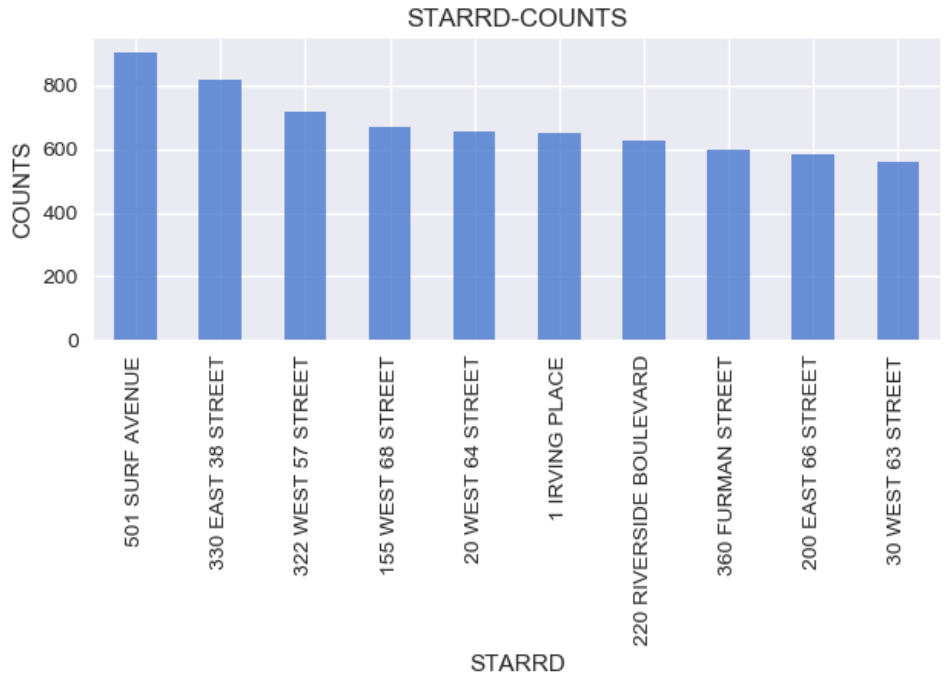
Field Name	*Description*
EXTOT	The exempt value of the total properties. Continuous data with metric.



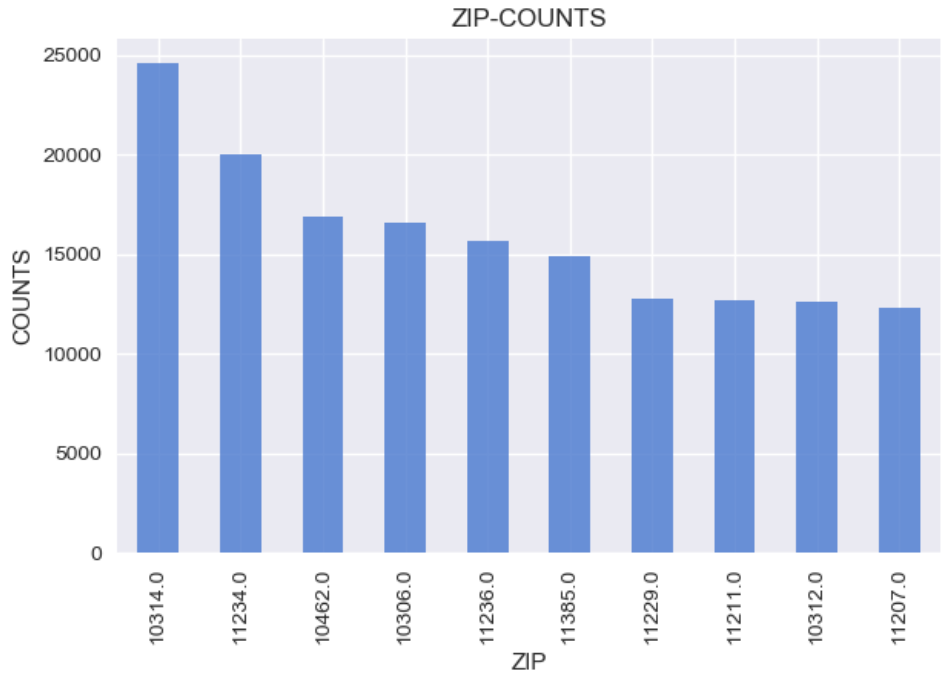
Field Name	*Description*
EXCD1	The first exempt condos. Continuous data with metric.



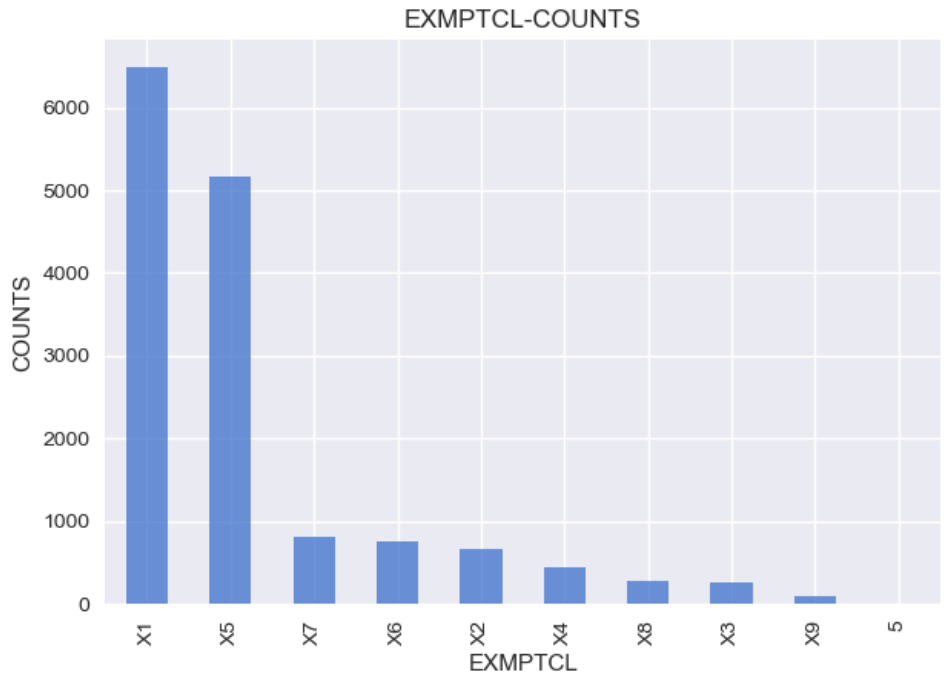
Field Name	*Description*
STADDR	Street address. The properties' specific street addresses. Characters without metric.



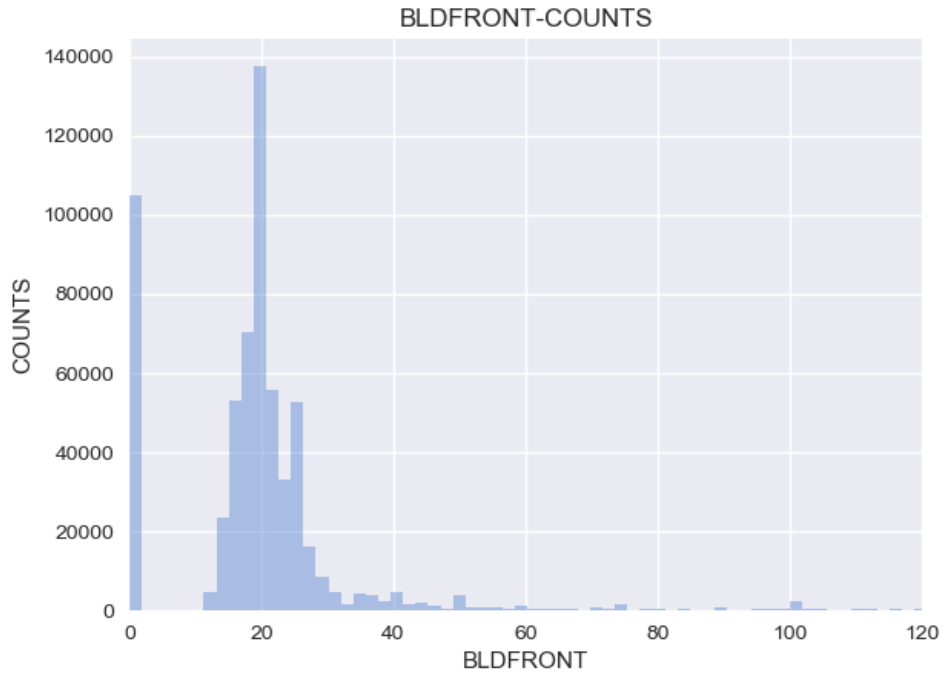
Field Name	*Description*
ZIP	The zip codes of the properties' locations. Characters without metric.



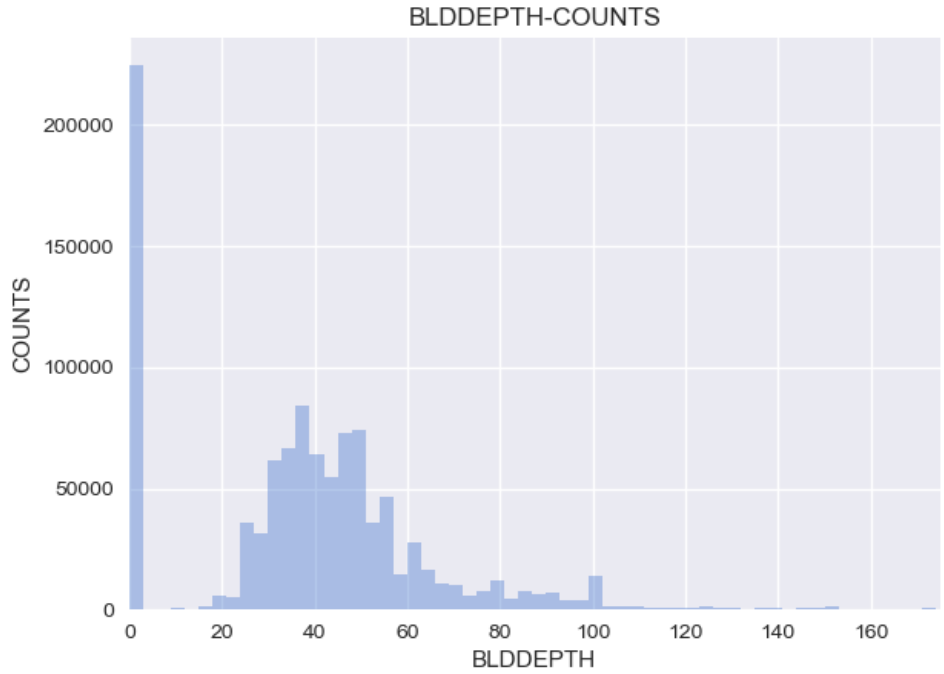
Field Name	*Description*
EXMPTCL	Exempt class for fully exempt properties. Categorical data without metric.



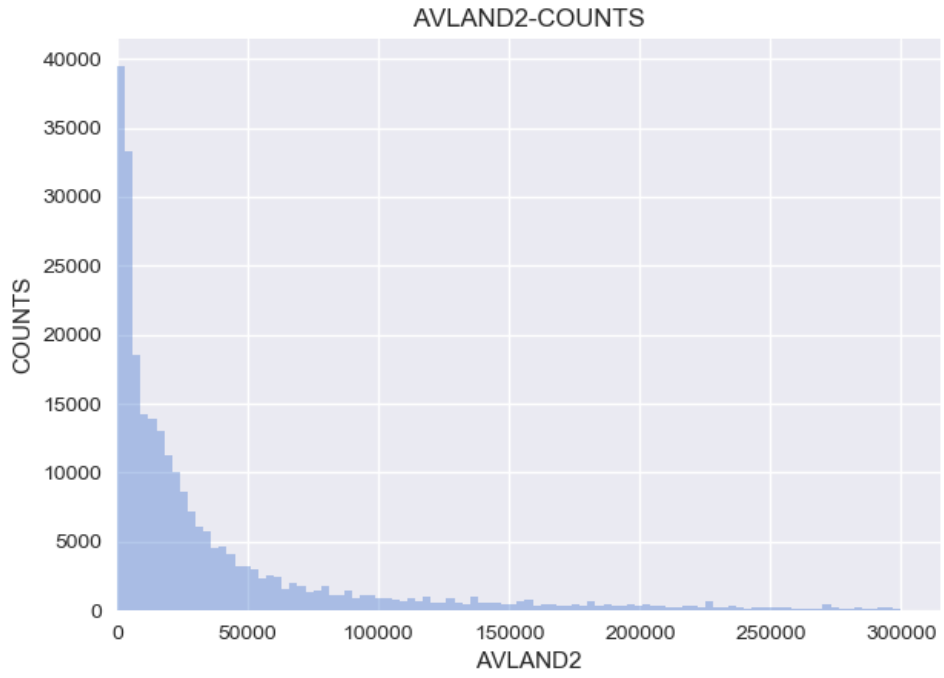
Field Name	*Description*
BLDFRONT	Building frontage, which means the length of the front lot line. Continuous data with metric.



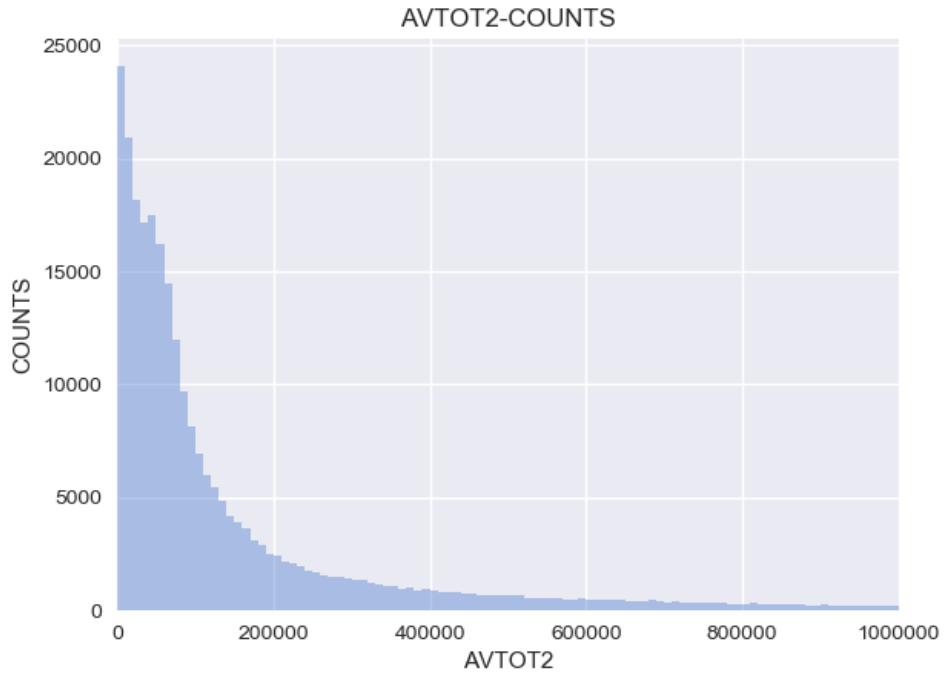
Field Name	*Description*
BLDDEPTH	Building depth means the horizontal distance between the front and rear property lines of a building, measured along a line midway between the side property lines. Continuous data with metric.



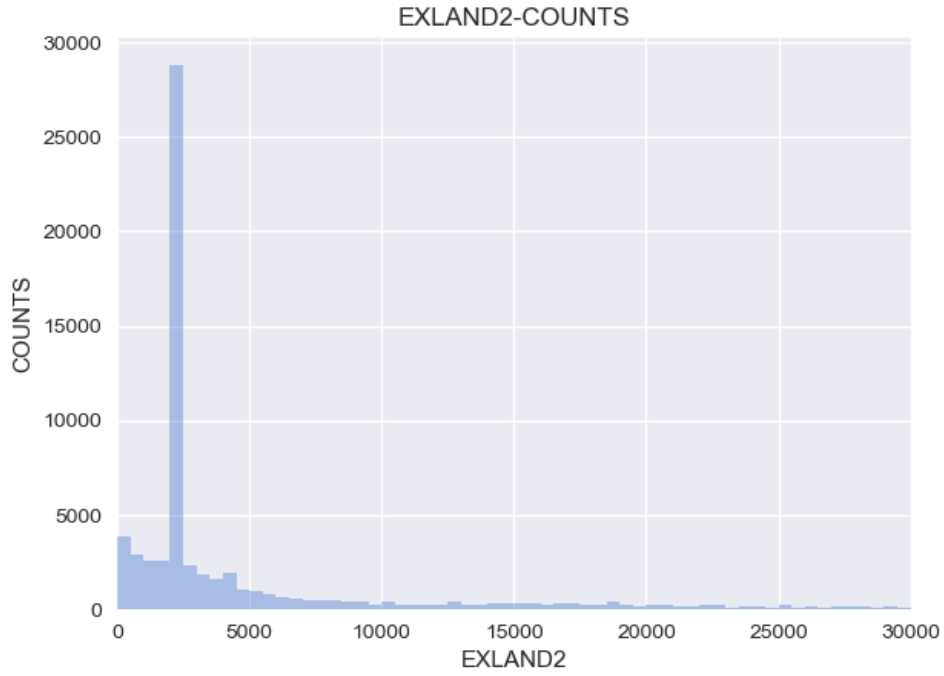
Field Name	*Description*
AVLAND2	The assessed value of the properties' lands. Continuous data with metric.



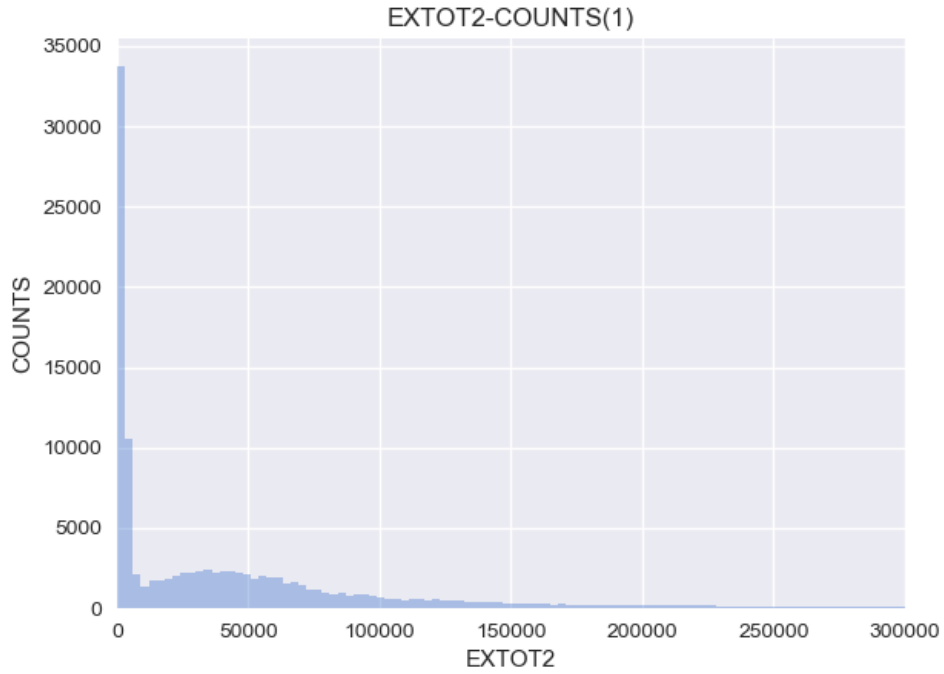
Field Name	*Description*
AVTOT2	The assessed value of the total properties. Continuous data with metric.

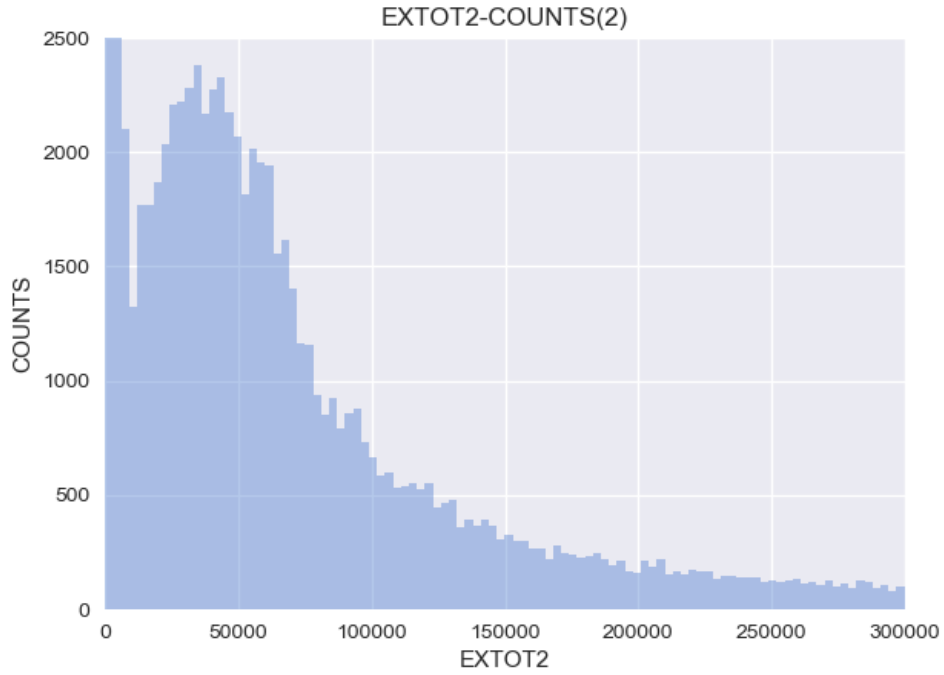


Field Name	*Description*
EXLAND2	The exempt value of the properties' lands. Continuous data with metric.

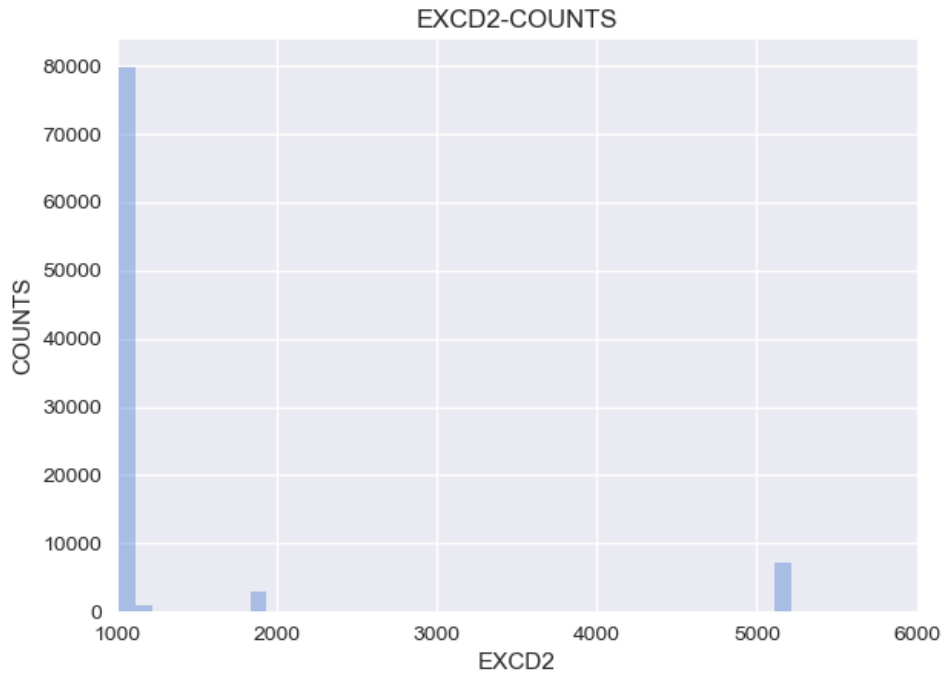


Field Name	*Description*
EXTOT2	The exempt value of the total properties. Continuous data with metric.

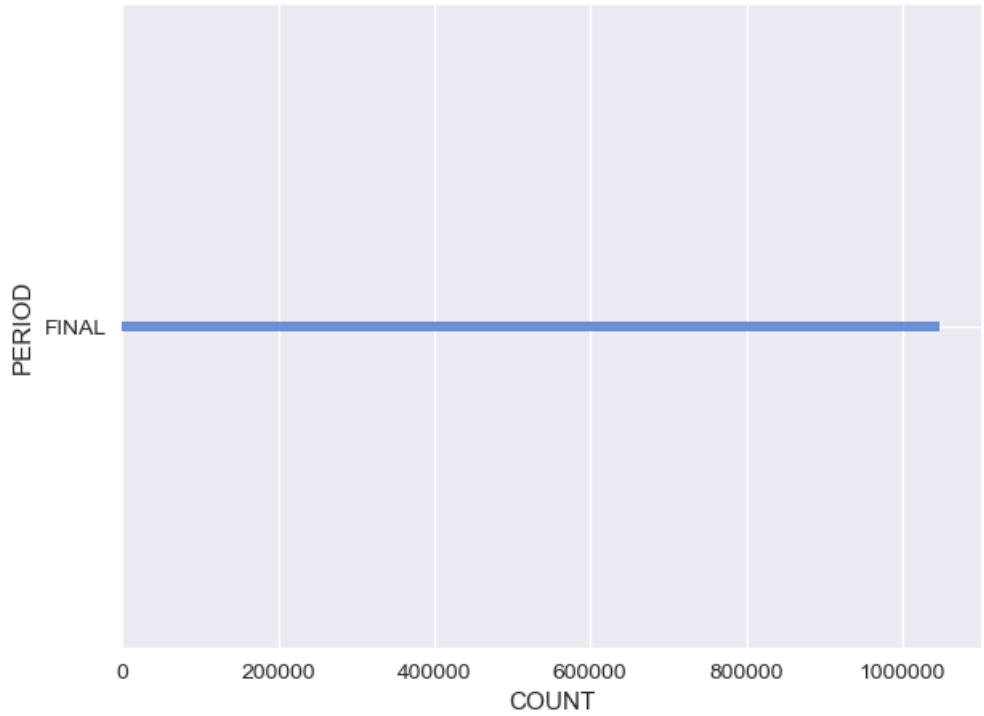




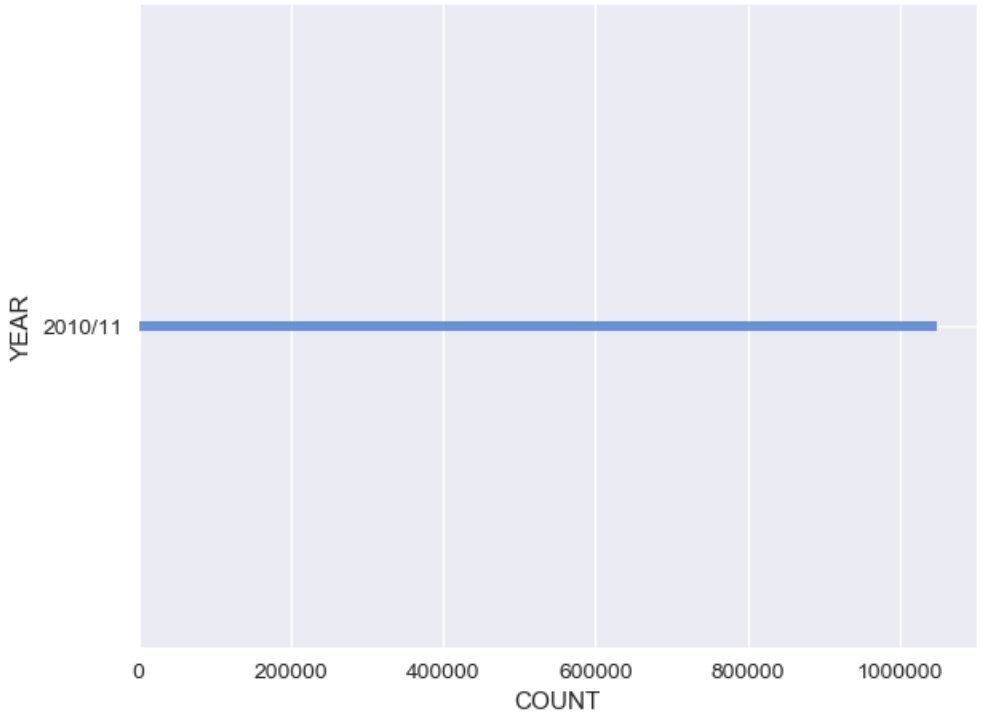
Field Name	*Description*
EXCD2	The second exempt condos. Continuous data with metric.



Field Name	*Description*
PERIOD	A characteristic of properties. Always be FINAL here. Constant.



Field Name	*Description*
YEAR	The year of the data source. Always be 2010/11 here. Constant.



Field Name	*Description*
VALTYPE	The value types. Always be AV-TR here. Constant.